Summer 2017

Lecture Notes for OpenStax Introductory Statistics

Daphne Skipper
United States Naval Academy

Neal Smith
Augusta University, nsmith12@augusta.edu

Robert Scott
Augusta University, rscott5@augusta.edu

Marvalisa Payne
Augusta University, mpayne@augusta.edu

Christopher Terry
Augusta University, cterry2@augusta.edu

Ancillary Materials Set
Augusta University

UNIVERSITY SYSTEM OF GEORGIA

Neal Smith, Christopher Terry, Marvalisa Payne, Robert Scott, Daphne Skipper

Lecture Notes for OpenStax Introductory Statistics

Affordable Learning Georgia

Georgia's Virtual Library GALILEO
An Initiative of the University System of Georgia

Lecture Notes for OpenStax Introductory Statistics

These are notes that can be used by either students or instructors in conjunction with the OpenStax Introductory Statistics textbook:

https://openstax.org/details/books/introductory-statistics

CHAPTER 1: SAMPLING AND DATA

Lecture Notes for Introductory Statistics [1]

Daphne Skipper, Augusta University (2016)

In the first chapter we are introduced to several very important statistical terms and concepts. Warning: Notice that in the previous sentence, there is no mention of formulas or calculations. **This is not a typical math class.** It is really more of a "critical thinking" course. It is essential to realize that you must approach the study of statistics much differently than you would approach a class like College Algebra or Calculus. Be sure to give yourself time to understand the concepts deeply. Spend time *thinking* about the concepts and definitions. Read the book and start the homework early enough that you have time to really understand each problem. Memorizing definitions and techniques will not guarantee successful completion of the course.

1. Statistics, Probability, and Key Terms

Before taking a statistics class, most people associate the term "statistics" with what a statistician would call **descriptive statistics**. We are using descriptive statistics when we use numbers or graphs to summarize a data set by calculating a mean or making a bar graph, for example. Descriptive statistics are essential to the field of statistics, but they will consume only a small portion of the course. The primary goal of the course is to begin a study of inferential statistics: drawing conclusions based on limited information (our data) and quantifying the accuracy of our conclusions. Confidence intervals and hypothesis tests are two of the tools of inferential statistics that we will explore. The tools of inferential statistics require probability, so our course will include a study of basic probability concepts. In the previous paragraph, we mention that our information is limited. What does that mean? To understand, we need the concepts of population and sample.

Example 1. Suppose a pharmaceutical company seeks FDA approval for a particular medication that has migraine headaches as a rare side effect. The FDA may require a scientifically valid estimate of the percentage of consumers of the medication who experience this side effect.

The pharmaceutical company cannot possible know the exact percentage of all eventual users of the medication who will experience migraines. That would require knowing information about every single user ever of the medication, even future users. In the context of this study, the collection of *all users of the medication* is the **population** of interest.

Since it is impossible to test every member of the population, the company must do the next best thing and test a representative **sample** of the population. In other words, the company must run an experiment in which some people are given the medication, then use the data from this sample to draw conclusions about the population.

To summarize: the process of inferring information about a *population* based on a representative *sample* is the practice of inferential statistics.

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

A number that describes an aspect of a population is called a **parameter**, such as the mean of the population data. A number that describes an aspect of a sample is a **statistic**, such as the mean of the sample data. Generally, parameters are unknown values that we use inferential statistics to estimate.

Example 2. Identify the parameter and statistic in the previous example.

Notice that the parameter is unknown (and can never be known in this case), but the statistic is known after running the experiment and collecting data. **The statistic is our best guess of the unknown parameter.**

A **variable** is a characteristic of interest about a subject. We use capital letters to name variables, such as X or Y. The answer to the question "what is the variable in this scenario?" is a verbal description of one piece of information we are collecting from each subject. The **data** (the singular form is **datum**) are the actual values that are collected about each subject.

Example 3. Continuing the same example, identify the variable, X, and the possible data values, x.

Example 4. Identify the population, sample, parameter, statistic, variable, and data for the following study: You wish to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday in your English class you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk. Collaborative Exercise 1.2, Illowsky & Dean.

2. Data, Sampling, and Variation

Most data is either categorical or quantitative. Data that sorts the subjects into categories is called **categorical** (or **qualitative**) **data**. Eye color is an example of a categorical variable with possible data values such as "blue" and "brown". Often there will be only two categories of interest, such as the last example when "yes" and "no" were the possible data values. Data that describes a numerical aspect of a subject is called **quantitative data**. Quantitative data can be further subdivided into data that represents a count, **discrete data**, and data that represents a measurement, **continuous data**. Data that represents a measurement can technically land anywhere on the number line, not necessarily on a whole number, even though we often round measurements to the nearest whole number.

Example 5. Identify the data type of each of the following variables: categorical, quantitative discrete, or quantitative continuous:

(1) the number of shoes owned by a subject
(2) the type of car a subject drives
(3) where a subject goes on vacation
(4) the distance from home to the nearest grocery store
(5) the number of classes a student takes per year
(6) the amount of money spent on textbooks in a semester by a student
(7) movie ratings
(8) the weight of a sumo wrestler
(9) the number of correct answers on a quiz
(10) the temperature of a cup of coffee

If we wish to use sample data to infer characteristics of a population, the sample data must be **representative** of the population; it should have the same characteristics as the population. The way a researcher selects subjects from the population is called the **sampling method**.

The most ideal sampling method is a **simple random sample**, in which every set of n subjects has the same probability of being selected. This is like putting names in a hat and randomly selecting 5 names, for example. All of the inferential statistics methods we will study require that the data arise from a simple random sample, or SRS. The reason for this requirement is that if the data are not representative of the population, then the conclusions will be meaningless (GIGO: Garbage In, Garbage Out).

Often a simple random sample is not feasible, or at least not practical, so researchers will do their best to use other sampling methods that are likely to result in a representative sample. A few different sampling methods that may be used successfully are:

- **stratified sampling:** subjects are categorized by similar traits, then sample subjects are randomly selected from each category in numbers that are porportional to their numbers in the population. Example: 4 girls are randomly selected and then 6 boys are randomly selected from a population that is 40% female. Stratified sampling guarantees a representative sample relative to the categories that are used.
- **cluster sampling:** there is already a natural categorization of subjects, usually by location. A sample of *categories* is selected randomly, and *every subject* in each of the selected categories is part of the sample. Example: randomly select 10 elementary schools in Georgia, then sample every teacher at each of those schools. Cluster sampling is done for the sake of saving time and/or money.
- **systematic sampling:** sample every nth subject. Example: sample every 1000th m&m to weigh and measure. Systematic sampling is common in manufacturing.

Sampling bias occurs when some members of the population are more likely to be chosen than others, resulting in a sample that is not representative of the population. The following sampling methods are common causes of sampling bias:

- **convenience sample:** collect data in a way that is convenient for the researcher, with little regard for obtaining a representative sample. Example: a teacher sampling the students in her own classes to estimate the percentage of biology majors in the school.
- **voluntary response/self selection:** each subject actively chooses to participate. Example: an internet survey posted on a website. This could also be less obvious. Example: an observational study of the health benefits of drinking wheatgrass juice.

Example 6. Suppose you are a reporter for the school newspaper. The university just announced an institutional name change and you would like to write a story that includes an analysis of student opinion about the change.

(1) Identify the population of interest.
(2) Describe a sampling design of each type below that should result in a representative sample.
   (a) simple random sample
   (b) stratefied sample
   (c) cluster sample
   (d) systematic sample
(3) Describe a sampling design of each type below. Do you think it is likely to result in a representative sample of the population for this particular question?

(a) convenience sample
(b) voluntary response sample

It should be clear that there will always be natural variability in sampling. Two researchers could each select a sample of size 20 via a simple random sample from the same population, but they will end up with two different data sets; we would not expect the means of the two samples to match, nor would we expect either sample mean to match the population mean exactly. This natural variability is called **sampling error** and it is a natural, unavoidable aspect of sampling and inferential statistics. We can, however, manage sampling error to a degree: larger sample sizes result in smaller sampling errors. This is a common theme we will see throughout the course.

On the other hand, **nonsampling error** arises from factors that can and should be avoided. The following are examples of nonsampling errors that can arise in statistical studies:

- **sampling bias:** described above
- **self-interest study:** sometimes the researcher or organization funding research has personal interest (usually monetary or political) in the outcome of a study. It is important to be aware of self-interest studies and carefully assess the statistical methods used in those cases. Example: Dr. Andrew Wakefield's research involving the MMR vaccine and autism.
- **correlation/causality:** if we identify a correlation between two variables, it is not necessarily the case that one causes the other. Example: Ice cream causes drowning.
- **non-response:** a subject can always refuse to participate in a study, so this can become a problem if too many subjects refuse to participate. Example: phone survey when everyone has caller id
- **misleading use of data:** incorrect graphs, incomplete data, or lack of context. Could be intentional or unintentional.
- **undue influence/leading questions:** asking questions in ways that lead the subject to a particular response. Example: Do you agree with the administrations insightful decision regarding the name change?
- **confounding:** multiple factors could contribute to a particular outcome. Example: when testing a particular diet, all subjects are required to exercise regularly. (Was it the diet or the exercise that caused the weight-loss?)

3. Organizing Categorical and Quantitative Data

(The organization of categorical data appears in Section 1.2 of the textbook. Histograms are introduced in Chapter 2.)

When we collect **categorical data** from each subject, the outcome is simply a list of the categories with **frequencies** (the number of subjects in each category) or **relative frequencies** (the *percentage* of subjects in each category). For example, suppose a fast food company samples teenagers to decide on a new milkshake flavor. Even though the data is categorical, the final result is a list of categories and frequencies (counts): vanilla: 20, chocolate: 27, cookies and cream: 35. CAUTION: even though there is a count involved, the actual data (milkshake flavors) is categorical.

A **bar graph** is usually the best choice for displaying categorical data because bar graphs are very clear for comparing frequencies of different categories. The easiest of all bar graphs to read is the **Pareto chart**, in which the bars are sorted from tallest to shortest. Note that it only makes sense to sort the bars in this way if the categories do not already have a natural ordering. **Pie charts** are very popular

for displaying categorical data, though it is often difficult to compare similarly sized pie slices at a glance. Notice that with a pie chart, the slices must add up to 100%. The following charts (from Section 1.2 of the textbook) illustrate three different graphical representations of the same data.

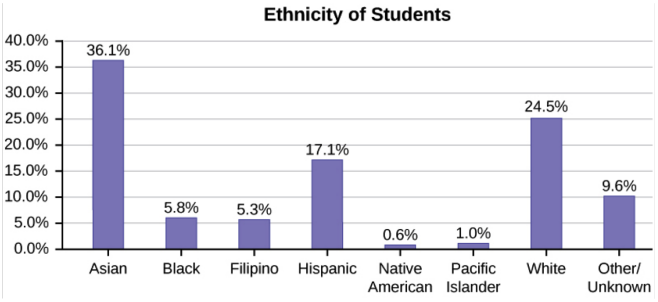

FIGURE 1. Relative frequency bar graph.

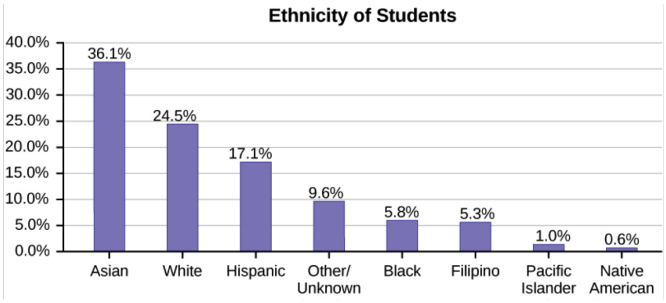

FIGURE 2. In a Pareto Chart, the categories are sorted by size.

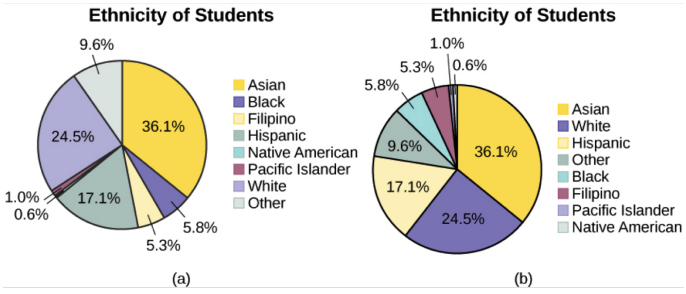

FIGURE 3. Pie charts.

We would like to organize **quantitative data** in a similar way, but the categories are not immediately available. Instead, we must split the range of data values into categories called **classes**. Then we can sort the data into these classes, treating the data much like we treated categorical data, finding a frequency and relative frequency for each class.

Example 7. The heights in inches of the students in an Elementary Statistics class are as follows:

64	64	65	75	71	61	64	73	64	66
73	67	67	67	63	63	60	66	62	74
74	67	65	71	67	65	67	72		

Complete the following frequency and relative frequency table. The last column, the **cumulative relative frequency**, is the percentage of data values that are no larger than the upper end of a given interval. (Later we will see that cumulative relative frequency is a **percentile**.)

Height in Inches	Frequency	Relative Frequency	Cumulative Relative Frequency
$55 - 59$			
$60 - 64$			

(1) What percentage of students in the class are 6" (5 feet,4 inches) or shorter?
(2) What percentage of students in the class are taller than 69" (5 feet, 9 inches)?

The next logical step is to draw a **histogram** of the data, which is a fancy word for a bar graph for quantitative data that has been sorted into classes. Histograms are the single most important graph in statistics, and we will be using them often. An important feature of a histogram is that there is a natural order for the classes (from left to right on the number line), so we will never sort a histograms bars by height.

Example 8. Draw a relative frequency histogram of the class height data from Example 7.

Example 9. CALCULATOR HISTOGRAM. Use the height data from Example 7.

(1) Clear calculator RAM.
(2) Enter the height into L_1.
(3) Use STAT PLOT to draw a histogram using default class boundaries. Use TRACE to see the default class boundaries and frequencies.
(4) Use WINDOW then GRAPH to change the histogram so that it uses the class boundaries we created in Example 7.

The reason histograms are so important is because they display the **distribution** of the data: where the data values fall along the range of possible values. They help us answer important questions such as: Is the data spread out or mostly concentrated in a small region? Is the data symmetrically distributed? Is the data skewed left (a few extremely low data values) or skewed right (a few extremely high data values)?

Example 10. What does the histogram of the height data tell us about the heights of the people in the sampled statistics class? Sketch a histogram shape for each class described below.

(1) The class has a lot of tall people.
(2) Everyone in the class is 64".
(3) The students make a nice staircase shape when lined up from shortest to tallest.

4. Experimental Design and Ethics.

A statistical study can take the form of an observational study or an experiement. In an **observational study** the researcher collects data but does not introduce any change to the subjects. A political poll is an example of an observational study. In an **experiment** the researcher applies treatments to subjects, then observes the effects of the treatments.

The purpose of an experiment is to study the relationship between two variables. The researcher controls the **explanatory variable**, then measures the change in the **response variable**. The different values of the explanatory variable are called **treatments**. An **experimental unit** is one subject being tested.

It is often the case that the expectation of a treatment has a psychological effect that confounds the experiment: was it the treatment or the expectation of the treatment that affected the response variable? To manage this problem, researchers set aside a group of subjects to be the control group. The **control group** receives the **placebo** treatment, a treatment that cannot affect the response variable. It is vital that study participants not know if they are receiving the placebo or an effective treatment. It is called **blinding** when the subjects don't know which treatment (which could be a placebo) they are receiving. A **double-blind experiment** is when both the subjects and the researchers involved with the subjects are blinded.

Example 11. Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men are recruited as participants. The men are divided randomly into two groups: one group will take aspirin and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know if it is aspirin or a placebo. At the end of the study, researchers count how many men in each group have had heart attacks. Identify each the following for this scenario:

(1) population
(2) sample
(3) experimental units
(4) explanatory variable
(5) treatments
(6) response variable

Example 1.19, Illowsky & Dean.

Ethics comes into play in statistical studies in a variety of ways. We have already seen that the self-interest studies can inspire fraudulent data and analysis. Another very important ethical component is related to the treatment of human subjects. There are laws that require that studies are safe, that participants understand the risks associated with a study, that subjects freely decide whether or not to participate, and that each subjects privacy is protected. Research institutions (including Augusta University) have **Institutional Review Boards (IRB)** to oversee the research at the institution and ensure the safety of all human subjects.

CHAPTER 2: DESCRIPTIVE STATISTICS

Lecture Notes for Introductory Statistics[1]

Daphne Skipper, Augusta University (2016)

1. Stem-and-Leaf Graphs, Line Graphs, and Bar Graphs

The **distribution** of data is how the data is spread or distributed over the range of the data values. This is one of the first and most important aspects one would want to know about any data set.

Stem-and-leaf graphs provide a quick way to view the distribution of a small data set by hand.

Example 1. Stem-and-leaf. The following are the ages of 29 actors at the time that they won the Best Actor award:

18	21	22	25	26	27	29	30	31	33
36	37	41	42	47	52	55	57	58	62
64	67	69	71	72	73	74	76	77	

Data from Example 2.18 in the textbook.
Explore the distribution of the best actor data by constructing a stem-and-leaf graph.

We already saw *bar graphs* in Chapter 1. Frequency polygons and time series graphs are two types of *line graphs* that we will see in the next section.

2. Histograms, Frequency Polygons, and Time Series Graphs

A **histogram** is similar to a bar graph. The difference is that bar graphs are for categorical data, whereas histograms are for quantitative data. Since quantitative data aren't naturally sorted into categories, we must create "classes" of data: equal width intervals of data values. We record classes and frequencies in a frequency table. We constructed a frequency table and histogram using height data in Chapter 1.

A **frequency polygon** displays the same data as a histogram, but in line graph form. This format is useful for comparing the distributions of multiple datasets by overlaying frequency polygons.

Example 2. Overlaying frequency polygons. Figure 1 is an overlay of frequency polygons of men's and women's pulse data. By overlaying frequency polygons, we are able to easily compare the distributions (histograms) of two data sets. (This chart is from Elementary Statistics by Mario Triola.)

A **time series graph** displays the trend of a variable over time. The x-axis is time (years, minutes, seconds, etc.). The y-axis is the range of data values.

Example 3. Time series graph. The time series graph in Figure 2 allows us to easily see the trends in housing prices in Australia versus the US.

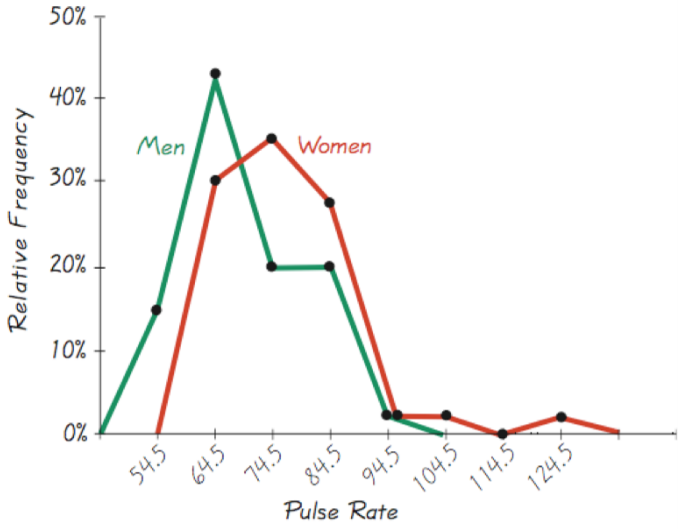

FIGURE 1. Frequency Polygons

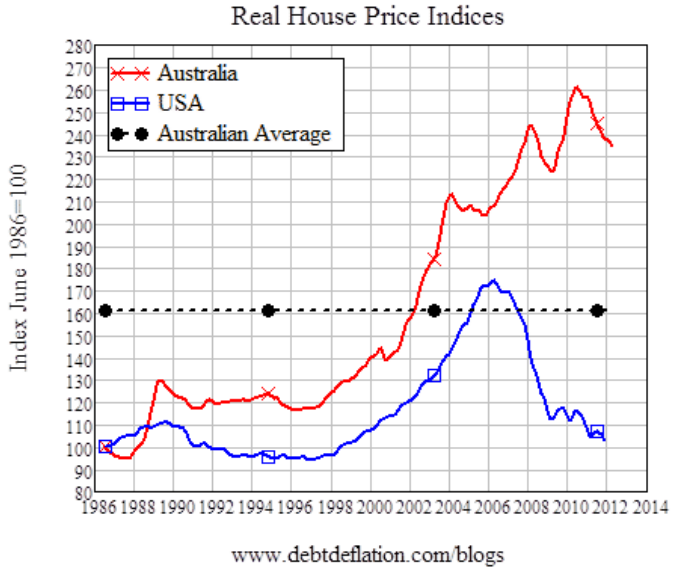

FIGURE 2. Time Series Graph

This TED talk contains an amazing example of data visualization: `http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen`. (Start at around the 2:30 mark.) Notice how Hans Rosling's methods make it easy to see patterns in extremely complicated time-series data.

[1]These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

3. Measures of Location of Data

The **median** of a data set is the "middle value", but it does not have to be one of the data values.

Example 4. Median. Find the median.

(1) 5 5 6 8 12

(2) 5 5 6 8 12 17

The **k^{th} percentile**, P_k, of a data set is the value that separates the lower $k\%$ of data values from the upper $(100-k)\%$ of data values.

Example 5. Understanding percentiles.

(1) The 35^{th} percentile, or P_{35}, separates the lower _____% of data values from the upper _____% of data values.

(2) "Median" is another name for the _____th percentile.

Percentiles are calculated on data that is sorted from lowest to highest. Low percentiles correspond to low data values. High percentiles correspond to high data values.

Example 6. Interpreting percentiles.

(1) Joe ran a 5K and his finishing time is the 5^{th} percentile. Interpret the 5^{th} percentile in this context. (What percent of racers did Joe beat? Is this a good finishing time?)

(2) If Abby takes the SAT, would she prefer for her score to be P_{10} or P_{88} of SAT scores for this year? Explain.

Locating the k^{th} percentile. The *position* in the sorted data set of the k^{th} percentile is

$$i = \frac{k}{100}(n+1).$$

If i is a whole number, P_k is the data value found at position i of the sorted data set. If i is a decimal, round i up and round i down; P_k is the average of the data values found at these two positions in the sorted data set.

Example 7. Calculating percentiles. The following are the sorted heights in inches of 40 students in a statistics class.

61	61	62	62	63	63	63	65	65	65
66	66	66	66	66	67	67	67	68	68
68	68	68	68	68	68	69	69	69	69
69	69	70	71	72	72	72	73	73	74

Data from TryIt 2.24 in the textbook.

(1) Find the 80^{th} percentile of the class heights. Use appropriate notation to express the answer. Interpret the answer.

(2) Find the 38^{th} percentile of the class heights. Use appropriate notation.

Finding the percentile of a data value. let x be the number of data values *below* that data value and let y be the number of data values *equal to* that data

value in the data set. The the percentile k of the data value is

$$k = \frac{x + 0.5y}{n}(100).$$

Example 8. FINDING THE PERCENTILE OF A DATA VALUE. Use the class height data from Example 7. Use appropriate notation to express your answer.

(1) Kara is 67 inches tall. At what percentile is Kara's height?
(2) Charles is 73 inches tall. At what percentile is Charles' height?

4. BOXPLOTS

Quartiles are numbers that separate the data into quarters. Like all percentiles, quartiles may or may not be actual data values. The first quartile, Q_1, is the middle of the bottom half of data; $Q_1 = P_{25}$. The second quartile, Q_2, is the median; $Q_2 = P_{50}$ = median. The third quartile, Q_3 is the middle of the top half of the data; $Q_3 = P_{75}$. The **five number summary** of a data set is: minimum, Q_1, median, Q_3, maximum.

We can calculate quartiles "by hand" just as we calculate any other percentiles. However, the calculator will calculate quartiles for us directly.

Example 9. USING THE CALCULATOR TO FIND QUARTILES. The following are the heights in inches of 20 boys in a statistics class.

66	66	67	67	68	68	68	68	68	69
69	69	70	71	72	72	72	73	73	74

The following are the heights of the 20 girls in the same statistics class.

61	61	62	62	63	63	63	65	65	65
66	66	66	67	68	68	68	69	69	69

Data from TryIt 2.24 in the textbook.

(1) Enter the 20 boys' heights into L_1.
(2) Find the 5 number summary for the boys' heights. (Use STAT, CALC, 1 VAR STATS.)
(3) Enter the 20 girls' heights into L_2.
(4) Find the 5 number summary for the girls' heights.

A **boxplot** is a graph of the 5 number summary scaled on a numberline, showing the concentration of data. The spread of the middle 50% of data values, the data values between Q_1 and Q_3, are indicated by a box. The median is marked in the box. The bottom 25% and top 25% of data values are indicated by the "whiskers".

Example 10. CONSTRUCTING A BOXPLOT. Use boys' and girls' height data from Example 9.

(1) Construct a boxplot for the boys' height data by hand.
(2) Use the calculator to construct boxplots for both data sets on the same scale.

Example 11. INTERPRETING BOXPLOTS. Use the boxplots from Example 10.

(1) 25% of girls are *shorter* than _____ inches. What quartile is this?
(2) 50% of girls are *shorter* than _____ inches. What quartile is this?

(3) 25% of boys are *taller* than _____ inches. What quartile is this?
(4) In which quartile are the girls' heights most concentrated? The boys' heights?

The **interquartile range**, or **IQR**, is the range of the middle 50% of data values: $\text{IQR} = Q_3 - Q_1$. The IQR is often used to identify outliers in the following way. Data values that are below $Q_1 - (1.5)\text{IQR}$ or above $Q_3 + (1.5)\text{IQR}$ are considered outliers.

Example 12. Identifiying outliers. Use the girls' class height data from the last few examples.

(1) Calculate the IQR of the girls height data.
(2) Find the boundary height below which a girls' height would be considered a "short" outlier.
(3) Find the boundary height above which a girls' height would be considered a "tall" outlier.
(4) According to these boundary values, does this data set contain any outliers?

5. Measures of the Center of Data

The **mean** is the most commonly used measure of the "center" of data. The mean is calculatated by adding all the data values and dividing by n, sample size, or the number of data values. Notation for mean:

$$\begin{aligned} \bar{x} &= \text{sample mean,} \\ \mu &= \text{population mean.} \end{aligned}$$

The **median** is the second most commonly used measure of the center. Since the median does not take the actual exact values of data into account, it is a better choice when there are extreme data values. In this case, the mean can be skewed toward the extreme data value(s) and be misleading with respect to the center of the bulk of the data. For example, we hear about "median income" rather than mean income, because there are some extremely large incomes that would make the ordinary income, as measured by the mean, appear larger than it really is for the average person.

The **mode** is the most frequently occuring data value. There can be more than one mode if there is a tie for the data value with the highest frequency. The mode is used primarily for categorical data, for which the mean and median don't make sense.

Example 13. Mode. Find the mode of the boys' height data:

66	66	67	67	68	68	68	68	68	69
69	69	70	71	72	72	72	73	73	74

Example 14. Mean using a frequency table. Use the same boys' height data as above.

(1) Make a frequency table of the height data.
(2) Use the frequency table to calculate the mean of the height data by hand.

(3) Using the calculator, enter the frequency table into L_1 and L_2.
(4) Calculate the mean and median using 1 Var Stats.

We can't get the exact mean of data from a *grouped* frequency table, but we can estimate the mean by using the midpoint of each class in place of each data value in that class.

Example 15. Mean of a grouped frequency table. Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data.

Hours spend on video games	Number of teens
0-3	3
4-7	7
8-11	12
12-15	7
16-19	9

(1) Find the midpoint of each class.
(2) What is the best estimate of the mean number of hours spent playing video games?

TryIt 2.30.

The **Law of Large Numbers** says that as sample size increases, sample mean ($\bar{x}$) gets closer and closer to the population mean (μ).

6. Skewness and the Mean, Median, and Mode

Data is skewed to the left if there is a "tail" of data to the left in the histogram, or on the low end of the data. Data is skewed to the right if there is a "tail" of data to the right in the histogram, or on the high end of the data.

The mean will be pulled in the direction of the skewing due to extreme data values. If there is more than mild skewing, the median is a more approprate measure of center than the mean for a more accurate summary of the bulk of the data.

7. Measures of the Spread of Data

Standard deviation is the most commonly used measure of the spread of data. The standard deviation tells us *approximately* how far data values are from the mean, on average. A larger standard deviation indicates that the data values are farther from the mean, on average.

Example 16. Interpreting standard deviation. Suppose the average (mean) wait time in line at both Publix and BiLo is 5 minutes. However, the standard deviation of the wait times at Publix is 1 minute and the standard deviation of the wait times in BiLo is 3 minutes.

(1) Which supermarket has more variation in wait times?
(2) At which supermarket would you be able to predict your wait time more precisely?

The standard deviation can be used to determine if a data value is close to or far away from the mean relative to the rest of the data. As a rule of thumb, more than

two standard deviations from the mean is considered "unusual". (Unusual data values are not as extreme as "outliers." More than three standard deviations from the mean is a more reasonable boundary for outliers, if we want to use standard deviation instead of the IQR formula.)

Example 17. COMPARING DATA VALUES USING STANDARD DEVIATION. Suppose the wait times at Kroger are 5 minutes with a standard deviation of 2 minutes.

(1) Draw a number line and mark the mean at 5 minutes and mark 1, 2, and 3 standard deviations above and below the mean. Label the unusual data value range(s) on the number line.
(2) Rosa waits 3 minutes at Kroger. How many standard deviations from the mean is her wait time?
(3) Binh waits 11 minutes at Kroger. How many standard deviations from the mean is his wait time?
(4) Is either wait time unusual at Kroger?

The standard deviation formulas are a bit much to calculate by hand, so we will use the calculator to find standard deviation. The formulas for sample and population standard deviation are different and your calculator provides both, so it is very important to be able to identify the notation for each when using the calculator:

$$\begin{aligned} s &= \text{sample standard deviation,} \\ \sigma &= \text{population standard deviation.} \end{aligned}$$

The calculator uses subscripts on these symbols to indicate which variable is being used: s_x and σ_x, for example. **We will almost always have sample data (as opposed to population data), so we will almost always use s_x from the calculator for standard deviaton.**

Example 18. STANDARD DEVIATION USING CALCULATOR. Use your calculator to find the standard deviation of the boys' height data:

66	66	67	67	68	68	68	68	68	69
69	69	70	71	72	72	72	73	73	74

Consider the data to be a sample from the population of male students at the school. (Is it reasonable to consider this to be a representative sample?) You may wish to enter the data in frequency table format, if you don't still have it in your calculator.

The **deviation** of a data value from the mean is the *signed* distance of the data value from the mean:

$$\text{deviation} = \text{data value} - \text{mean}.$$

The **z score** of a data value is the number of standard deviations from the mean that data value is. If the data values is below the mean, the z score is negative. If

the data value is above the mean, the z score is positive.

$$\begin{aligned} z &= \frac{\text{deviation}}{\text{standard deviation}} \\ &= \frac{\text{data value} - \text{mean}}{\text{standard deviation}} \\ &= \frac{x - \bar{x}}{s} \end{aligned}$$

Example 19. NOTATION. Write the formula for a *population* z score, using the appropriate population symbols.

Example 20. Z SCORE. Use the same boys' height data, which has mean $\bar{x} = 69.5$ and standard deviation $s = 2.5$.

(1) Find the z score of a height of 74 inches. Should we consider 74 inches to be unusually tall among these boys?
(2) Find the height that has a z score of -0.7.

Example 21. COMPARE DATA VALUES FROM DIFFERENT DATA SETS. Find the z score for each girl's time relative to her team. Which girl has the fastest time relative to her teammates?

Swimmer	Time(sec)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

TryIt 2.35

The following two rules give us more precise ideas of how far the data are spread from the mean based on the standard deviation. Chebyshev's Rule is for ANY data set:

(1) At least 75% of data is within two standard deviations of the mean
(2) At least 89% of the data is within three standard deviations of the mean
(3) At least 95% of the data is within 4.5 standard deviations of the mean

The Empirical Rule, also known as the **68-95-99 Rule**, is ONLY for data with a BELL-SHAPED and SYMMETRIC histogram/distribution:

(1) Approximately 68% of data is within one s.d. of the mean
(2) Approximately 95% of data is within two s.d.s of the mean
(3) Approximately 99% of data is within one s.d.s of the mean

The formula for sample standard deviation is

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}.$$

The formula for population standard deviation is

$$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}.$$

We won't use these formulas in practice, but it would be good to get a feel for how the formula approximates the average distance of data values from the mean, and to recognize that the formulas for *sample* standard deviation and *population* standard deviation differ.

Example 22. THE STANDARD DEVIATION FORMULA. In groups of 3 or 4 students, guess the age of the person in the provided photograph. Submit your group's guess. *Instructor note: Provide a photograph of someone that you know, but who is unknown to the students for this short activity. It's also fun with more photographs.*

(1) List each guess in the first column a table.
(2) In the second column, list the deviation of each guess from the true age: guess - true age.
(3) Does a negative deviation indicate a low or high guess?
(4) Are the guesses biased young? Biased old? Relatively unbiased?
(5) In the third column, list the absolute value of the deviation from the previous column. The most natural average distance from the true age is the *mean absolute deviation*. However, absolute values are not practical in algebraic formulas. Calculate the mean absolute deviation. This is the average distance of the guesses from the true age.
(6) In the fourth column, list the square of each deviation.
(7) Now calculate the variance (the square of the standard deviation): Add up the values in the 4th column and divide by the $n-1$. The units of variance is "years squared".
(8) Take the square root of the variance. This is the standard deviation of the guesses from the true age. The units of standard deviation is "years", the same as the data values.
(9) Compare the mean absolute deviation and the standard deviation. (Are they the same?)
(10) Find the z score of your group's guess. Which group had the best guess? Did any group have an "unusual" guess?

Adapted from "Teaching Statistics, a bag of tricks" by Gelman and Nolan.

guess	deviation (guess - true age)	|deviation|	$(\text{deviation})^2$

CHAPTER 3: PROBABILITY TOPICS

LECTURE NOTES FOR INTRODUCTORY STATISTICS [1]

Daphne Skipper, Augusta University (2016)

1. TERMINOLOGY

In this chapter, we are interested in the probability of a particular event occurring when we conduct an "**experiment**". The **sample space** of an experiment is the set of all possible outcomes of the experiment. An **event** is one or more outcomes of the experiment. A single outcome is called a **simple event**. Notation:

$$\begin{aligned} S &= \text{sample space} \\ A, B, C, ... &= \text{event} \\ P(A) &= \text{probability of event } A \end{aligned}$$

Example 1. TERMINOLOGY. Rolling a fair die is an **experiment**.

- The **sample space** of this experiment is $S = \{1, 2, 3, 4, 5, 6\}$. (Use set notation to list the outcomes of an event.)
- $A =$ "rolling an even number" is an example of an **event** consisting of three outcomes from the sample space: $A = \{2, 4, 6\}$.
- $B =$ "rolling a 3" is an example of a **simple event**: $B = \{3\}$.

Probabilities are numbers between zero and one. $P(A) = 1$ if event A is certain to occur. $P(B) = 0$ if event B can never happen. If all of the simple events in a sample space are *equally likely* to occur,

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S}.$$

The following are various ways of combining events:

A OR B = A ∪ B (union): Every outcome in A along with every outcome in B, including outcomes that are in both events.

A AND B = A ∩ B (intersection): Only outcomes that are in both A and B.

A′ (complement): Every outcome in S that is NOT in A.

A given B = A|B (conditional event): The event that A happens assuming that B has happened. This has the effect of *reducing the sample space from S to B.* So,

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\text{number of outcomes in } A \text{ AND } B}{\text{number of outcomes in B}}.$$

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

Example 2. Probability notation. The students at an elementary school have various characteristics that may be of interest to the school administration. Consider the following events:

$$\begin{aligned} S &= \text{a student eats school lunch,} \\ L &= \text{a student brings a lunch box,} \\ B &= \text{a student rides the bus,} \\ W &= \text{a student walks to school.} \end{aligned}$$

Write the symbols for the following probabilities. You may use OR and AND *or* $\cup$ and $\cap$.

(1) the probability that a student does not bring a lunch box
(2) the probability that a student rides the bus or eats school lunch
(3) the probability that a student brings a lunch box and walks to school
(4) the probablity that a student who rides the bus brings a lunch box

Example 3. Simple probabilities. A bowl contains 22 jelly beans: 6 red, 8 white, 3 yellow, and 5 green. Randomly select a single jelly bean from the bowl. Consider the following defined events:

$$\begin{aligned} R &= \text{the selected jelly bean is red,} \\ W &= \text{the selected jelly bean is white,} \\ Y &= \text{the selected jelly bean is yellow,} \\ G &= \text{the selected jelly bean is green.} \end{aligned}$$

Calculate the following probabilities

(1) $P(R)$
(2) $P(R')$
(3) $P(Y \text{ OR } G)$
(4) $P(W \text{ AND } Y)$
(5) $P(R|G)$

2. Independent and Mutually Exclusive Events

Two events A and B are **independent** if the knowledge that one occurs does not affect the probability that the other occurs. **Events A and B are independent if any of the following are true (we must check *only one*):**

$$\begin{aligned} P(A|B) &= P(A), \\ P(B|A) &= P(B), \\ P(A \text{ AND } B) &= P(A)P(B). \end{aligned}$$

Example 4. Identify independent events intuitively. Identify each pair of events as independent or dependent.

- the outcomes of two consecutive rolls of a fair die
- the gender of the first child born to each of two couples
- taking swim lessons and learning to swim

- from a deck of cards, randomly drawing two cards *with replacement*, and both are aces
- from a deck of cards, randomly drawing two cards *without replacement*, and both are aces

Example 5. IDENTIFY INDEPENDENT EVENTS BY VERIFYING EQUATION. Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are A and B independent? We must only check one equation in general, but this time check all three. TryIt 3.8.

Events A and B are **mutually exclusive** if they can not both happen at the same time. If one happens, then the other cannot happen. **Events A and B are mutually exclusive if and only if**

$$P(A \text{ AND } B) = 0.$$

Example 6. TREE DIAGRAM; IDENTIFY MUTUALLY EXCLUSIVE AND INDEPENDENT EVENTS. Consider the experiment of flipping a fair coin two times. Record heads (H) or tails (T) for each flip.

(1) Draw a **tree diagram** of the experiment.
(2) Use the tree diagram to list the sample space of the experiment.
(3) For each event, list the outcomes and find the probability.
   (a) A = at least one head
   (b) B = only tails
   (c) C = only heads
(4) Find $P(A \text{ AND } B)$, $P(A \text{ AND } C)$, and $P(B \text{ AND } C)$.
(5) Which pairs of events are mutually exclusive?
(6) Find $P(C|A)$.
(7) Are events C and A independent?

Variation on Example 3.9.

Example 7. IDENTIFY INDEPENDENT AND MUTUALLY EXCLUSIVE EVENTS. A student goes to the library. Let events B = "the student checks out a book" and D = "the student checks out a DVD." Suppose that $P(B) = 0.40$, $P(D) = 0.30$, and $P(B \text{ AND } D) = 0.20$.

(1) Find $P(B|D)$.
(2) Find $P(D|B)$.
(3) Are B and D independent?
(4) Are B and D mutually exclusive?

TryIt 3.10.

3. Two Basic Rules of Probability

The **multiplication rule** states that for events A and B defined on a sample space,

$$P(A \text{ AND } B) = P(A)P(B|A).$$

If A and B are *independent events*, then $P(B|A) = P(B)$ and the rule simpifies to

$$P(A \text{ AND } B) = P(A)P(B).$$

Notice that the multiplication rule is a rearrangement of the formula for the conditional probability $P(B|A)$.

Example 8. Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. Let

C = the event that she makes the first shot (P(C) = 0.75), and
D = the event that Helen makes the second shot (P(D) = 0.75).

However, Helen is more likely to make the second shot if she has made the first: the probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws? TryIt 3.15.

The **addition rule** states that for events A and B defined on a sample space,

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B).$$

If A and B are *mutually exclusive events*, then $P(A \text{ AND } B) = 0$ and the rule simplifies to

$$P(A \text{ OR } B) = P(A) + P(B).$$

Example 9. Continuing Example 8, find the probability that Helen makes the first or the second free throw (or both).

Example 10. A school has 200 seniors of whom 140 will be going to college (C) next year. Forty will be going directly to work (W). The remainder are taking a gap (G) year. Fifty of the seniors going to college play sports (S). Thirty of the seniors going to directly to work play sports. Five of the seniors taking a gap year play sports.

(1) What is the probability that a senior is going to college or directly to work? (Classify these events: independent? mutually exclusive?)
(2) What is the probability that a senior plays sports?
(3) What is the probability that a senior is going to college and plays sports? (Classify these events: independent? mutually exclusive?)
(4) What is the probability that a senior is going to college or plays sports?

Adapted from TryIt 3.16 and TryIt 3.18.

Example 11. A student goes to the library. Let events

B = the student checks out a book, and
D = the student checks out a DVD.

Suppose that

$$P(B) = 0.40,$$
$$P(D) = 0.30,$$
$$P(D|B) = 0.5$$

Describe and find the following probabilities:

(1) $P(B')$
(2) $P(D \text{ AND } B)$
(3) $P(D \text{ OR } B)$
(4) $P(B|D)$

TryIt 3.19.

4. Contingency Tables

A **contingency table** displays frequencies of data relative to two different variables. We will see contingency tables again in Chapter 11.

Example 12. Contingency table. Make a two-way contingency table for the information in Example 10. List row and column totals.

A school has 200 seniors of whom 140 will be going to college (C) next year. Forty will be going directly to work (W). The remainder are taking a gap (G) year. Fifty of the seniors going to college play sports (S). Thirty of the seniors going to directly to work play sports. Five of the seniors taking a gap year play sports.

Use your frequency table to find the following probabilities.

(1) What is the probability that a senior plays sports?
(2) What is the probability that a senior is going to college or directly to work? (Classify these events: independent? mutually exclusive?)
(3) What is the probability that a senior is going to college and plays sports? (Classify these events: independent? mutually exclusive?)
(4) What is the probability that a senior is going to college or plays sports?
(5) Express the probability in words and find the probability:
   (a) $P(S|W)$
   (b) $P(W|S)$

Adapted from TryIt 3.16.

5. Tree and Venn Diagrams

A **tree diagram** is a special type of graph that helps to determine the outcomes of experiments and to visualize and calculate probabilities.

Example 13. Tree diagram; independent events. Roll a die and flip a coin. Draw a tree diagram to show all the possible outcomes in the sample space. Label the edges with probabilites. Let the numbers 1, 2, 3, 4, 5, and 6, represent the event of landing on that number. Let E be the event of rolling an even number. Let H and T represent landing on heads and tails, respectively. Use the tree diagram to find the following probabilites.

(1) $P(5 \text{ AND } H)$
(2) $P(E)$
(3) $P(E \text{ AND } T)$
(4) $P(H|E)$
(5) $P(5')$

Probabilities on the edges of a tree diagram are particularly useful when the events are dependent. Note that when the events are dependent, every level after the first level has *conditional* probabilities on the edges.

Example 14. Tree diagram; dependent events. Suppose there are four green balls and 9 yellow balls in a box. Three balls are drawn from the box without replacement. Draw a tree diagram representing the experiment using probabilities on the edges. Use the tree diagram to find the probabilities of the following events.

(1) all three balls are green
(2) at least one ball is yellow
(3) one ball of each color is selected

Similar to Example 3.25

A **Venn diagram** is another way to organize the outcomes of an experiment graphically.

Example 15. VENN DIAGRAM. In a certain game, a fair 12-sided die is rolled. A player earns 2 points by landing on an even number (E) and 3 points by landing on a multiple of three (T).

(1) Draw a Venn diagram representing the sample space of the die roll.
(2) Express each of the following events in terms of E and T, then use the Venn diagram to find the probability of each.
   (a) winning 5 points on a single roll
   (b) winning at least 2 points on a single roll
   (c) going scoreless on a single roll

Example 16. VENN DIAGRAM; PROBABILITIES. In a bookstore, the probability that a customer buys a novel is 0.5, and the probability that the customer buys a non-fiction book is 0.3. Suppose that the probability that the customer buys both is 0.2.

(1) Draw a Venn diagram representing the situation. Use probabilities instead of outcome in each space.
(2) Find the probability of each event.
   (a) a customer buys either a novel or a non-fiction book
   (b) a customer does not buy a novel
(3) Customers who buy novels are entered into a drawing to win tickets to a book signing by a famous author. Add a circle to the Venn diagram representing the winners of the tickets.

Variation on TryIt 3.30.

CHAPTER 4: DISCRETE PROBABILITY DISTRIBUTIONS

LECTURE NOTES FOR INTRODUCTORY STATISTICS [1]

Daphne Skipper, Augusta University (2016)

A **random variable** is the the outcome of an experiment translated to a number. It is random in the sense that the particular value it takes depends on the outcome of the experiment. The verbal description of a random variable describes how to find or calculate the data value. In other words, the description of a particular random variable provides a recipe for turning every outcome of an experiment into a number. Random variables are named by capital letters, like X. The same letter but lowercase, like x, denotes a data value (*a number*).

Example 1. RANDOM VARIABLE.
Experiment: *Flip a coin four times.*
Random variable (DESCRIPTION): X = *the number of heads in those 4 flips.*
Data (NUMBERS): *The possible values of* x *are* $x =$ _____.

If you find the data value x by COUNTING, then X is called a **discrete random variable**. If you MEASURE to get the data value x, then X is a **continuous random variable**, because every value on a continuous interval of the number line is theoretically possible.

Example 2. Is the random variable X in Example 1 a discrete or continuous random variable? _____

We will study discrete random variables in this chapter. We will study continous random variables in chapters 5 and 6. Both are critical to the study of inferential statistics.

1. PROBABILITY DISTRIBUTION FUNCTION (PDF) FOR A DISCRETE RANDOM VARIABLE

A **probability distribution function, PDF,** for a discrete random variable assigns a probability to every single outcome (simple event) of the experiment. It should be no surprise that:

(1) each probability must be between _____ and _____, and
(2) if you add up all the probabilities of all the possible simple events, they must add up to _____.

We usually see a PDF for a discrete random variables expressed in the form of a **probability distribution table**. The first column should have all the possible values of x. The second column should tell the probability of getting each of those values.

Example 3. PDF. Consider the experiment of rolling a fair die one time and recording the number it lands on.

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

(1) Random variable $X =$ _____.
(2) All the values x can take are $x =$ _____.
(3) Make a **probability distribution table** for random variable X.
(4) Does the table satisfy the requirements for a PDF? In other words,
    (a) are all of the probabilities between 0 and 1?_____
    (b) do all of the probabilities of the simple events add up to 1?_____

Example 4. PDF. Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Randomly choose one week out of the semester.

(1) Random variable $X =$ _____.
(2) All the values x can take are $x =$ _____.
(3) Make a **probability distribution table** for random variable X.
(4) Does the table satisfy the requirements for a PDF?

Example 4.2

2. Mean (or Expected Value or Long Term Average) and Standard Deviation of a PDF

2.1. **Two Types of Tables.** You must have noticed that probability distribution tables look a lot like the relative frequency tables that we use to make bar graphs and histograms. Basically they are, with one big difference: relative frequency tables tell the percentage of times each outcome *actually occurred using real-live sample data values*, whereas probability distrubution tables provide the percentage of times you can *theoretically expect* each possible data value to occur.

So you could think of a relative frequency table as telling the history of an experiment, and a probability distribution table as predicting the future of an experiment.

Example 5. Probability distribution and relative frequency tables. In the dice example (Example 3) above, we made a *probability distribution table* for the experiment of rolling a die and recording the number the die landed on:

x	$P(x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

(1) Roll a die 20 times and make a *relative frequency table* of the outcomes.
(2) Which table did you use actual data to create?
(3) Which table did you use a theoretical understanding of dice to create?
(4) Are the proportions/percentages the same or different in each table? Why?
(5) If you rolled a die 1,000,000 times and made a relative frequency table, what would you expect the second column to look like?

The **Law of Large Numbers** says that when the sample size (number of times an experiment is run) gets really large, the proportions in the second column of a relative frequency table for the experiment get closer and closer to the theoretical proportions you would find in a probability distribution table for the experiment.

Example 6. Law of Large Numbers. The Law of Large Numbers is used to make probability distributions like the one in Example 4. We could use large amounts of historical sample data about Nancy's attendance habits (in the form of a relative frequency table) to make a theoretical probability distribution table to predict Nancy's future attendance habits.

2.2. **Expected Value and Standard Deviation of a PDF.** The **mean**, also called the **expected value** or the **long term average**, **of a probability distribution** is the number you would expect to get if you ran the experiment over and over MANY times, recorded each data value, then took the mean of all those data values.

The **standard deviation of a probability distribution** is the number you would expect if you ran the experiment over and over MANY times, recorded each data value, then took the standard deviation of all those data values.

We calculate the expected value (or long term average) and the standard deviation of a probability distribution in table form just like we do for a frequency table. **Notation:** Because these are theoretical values, not based on sample data, we use the parameter notation for mean and standard deviation, μ and σ, respectively.

Example 7. Expected Value and Standard Deviation of a Probability Distribution. The following table provides the *probability distribution* for the number of times a newborn baby's crying wakes its mother after midnight during the course of a week.

x	$P(x)$	$xP(x)$
0	$\frac{2}{50}$	
1	$\frac{11}{50}$	
2	$\frac{23}{50}$	
3	$\frac{9}{50}$	
4	$\frac{4}{50}$	
5	$\frac{1}{50}$	

(1) Random variable $X =$ _____
(2) According to the table, the possible data values for x are $x =$_____
(3) What is the most common number of times a mother is awakened?_____
(4) What is the least common number of times a mother is awakened?_____
(5) Find the expected value, or long-term average, of the number of times a newborn baby's crying wakes its mother after midnight. (The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight.) Use the appropriate notation. _____
(6) Calculate the standard deviation of this random variable. Use the appropriate notation._____
(7) Would it be unusual for a mother to be awakened four times? _____ Five times? _____

One really cool application of the expected value of a probability distribution is determining whether or not the odds are against you in a game of chance when money is at stake. In this case, we let the random variable X be the amount of money you win (or lose) when you play the game one time. (Use a "+" to indicate a win and a "-" to indicate a loss.)

Example 8. Expected Winnings of a Game of Chance. Suppose you play a game with a biased coin. You play each game by tossing the coin once. P(heads) $= \frac{2}{3}$ and P(tails) $= \frac{1}{3}$. If you toss a head, you pay \$6. If you toss a tail, you win \$10. If you play this game many times, will you come out ahead? To find out, complete the following:

(1) Random variable $X =$ _____
(2) The possible values x can take are $x =$ _____
(3) Make a probability distribution table for X.

	x	$P(x)$
WIN		
LOSE		

(4) Find the expected value of the probability distribution. (This is the amount you would expect to win per game on average over the long-run.) _____
(5) Is it a smart idea to play this game? _____

Example 4.6

3. Binomial Distribution

A lot is known about certain probability distributions. If we can classify an experiment as having one of these well-known probability distributions, we can take advantage of all of this prior knowledge.

The **binomial probability distribution** is a discrete probability distribution that often arises in situations involving *categorical data*.

3.1. **Recognizing the Binomial Distribution.** An experiment is a binomial experiment if it consists of repeated trials of a two outcome process. The **characteristics of a binomial experiment** are as follows:

(1) there are a set number of trials (n = number of trials or sample size),
(2) there are only two possible outcomes for each trial (called "success" and "failure"), and
(3) each trial of the experiment is independent and has the same probability of "success" ($p = P$(success) and $q = P$(failure) on a single trial)

The **binomial random variable** X = the number of "successes" out of n trials. The possible values x can take are $x =$ _____.

Notation: $X \sim B(n, p)$ means that "X is a random variable with a binomial distribution". The parameters are n = number of trials trials and p = probability of success on a single trial. If you know p, then $q = 1 - p$.

Example 9. Binomial random variable. Flip a coin 5 times. We are interested in the number of times the coin lands on heads.

(1) Random variable $X =$

(2) The possible values of x are $x =$
(3) Is X a binomial random variable? In other words:
  (a) Are there a set number of trials?
  (b) Are there two outcomes for each trial?
  (c) Are the trials independent with the same probabilities of the outcomes occurring for each trial? _____
(4) Identify
  (a) n
  (b) success
  (c) failure
  (d) p
  (e) q.
(5) $X \sim$

Important notes about Example 9 and binomial random variables in general:

- *Categorical data* is collected for each trial, for example "heads" or "tails."
- The binomial random variable X recorded at the end of all the trials is *quantitative discrete* because it COUNTS the number of "successes."
- The outcomes "success" and "failure" are words, not numbers.
- The outcome that is being counted is called a "success."
- p and q are probabilities that add up to 1. When you are figuring out p, think about what is happening on a *single trial*. When you find p, you can always get q by using $q = 1 - p$.
- Even though there are only two outcomes, it is rare that they have an equal probability of occurring. Usually $p \neq q$.

The binomial distribution often arises in scenarios involving surveys, like the following example.

Example 10. RECOGNIZING A BINOMIAL SCENARIO. According to an article in the Augusta Chronicle about the name change of Georgia Regents University to Augusta University, 74% of people in the U.S. correctly place Augusta in Georgia. Suppose we conduct a random sample of 20 people in the U.S., and we are interested in the number of those people who locate Augusta in Georgia.

(1) What is a single trial of the experiment?
(2) What are the possible outcomes of a single trial?
(3) Which of those outcomes is labeled "success"? "success" =
(4) How many trials are there? $n =$
(5) What are p and q in words?
(6) Find p and q.
(7) $X \sim$

3.2. **Expected Value and Standard Deviation of the Binomial Random Variable.** Both the mean and standard deviation of a binomial random variable have very simple formulas.

If X is a binomial random variable, $X \sim B(n, p)$, then the **expected value (mean)** of X is

$$\mu = np$$

and the **standard deviation** of X is

$$\sigma = \sqrt{npq}.$$

Example 11. Suppose $X \sim B(50, 0.2)$. Find the mean and standard deviation of X.

The expected value (mean) of a binomial random variable is the number of successes you would expect to get in n trials. We can use the standard deviation to determine "unusual" numbers of successes.

Example 12. According to an article in the Augusta Chronicle about the name change of Georgia Regents University to Augusta University, 74% of people in the U.S. correctly place Augusta in Georgia.

(1) Out of a random sample of 20 people in the U.S., about how many would we expect to say that Augusta is in Georgia?
(2) Find the standard deviation of the associated binomial random variable.
(3) Would it be unusual for only 7 of the people to say that Augusta is in Georgia?
(4) If only 7 say that Augusta is in Georgia, what would it make you think about the statement that 74% of peope in the U.S. place Augusta in Georgia?

3.3. **Finding binomial probabilities.** The following notation is useful:

$P(x = 4)$ = the probability that x is 4,
$P(x \leq 4)$ = the probability that x is "4 or less" or "at most 4,"
$P(x \geq 4)$ = the probability that x is "4 or more" or "at least 4,"
$P(x > 4)$ = the probability that x is greater than 4,
$P(x < 4)$ = the probability that x is less than 4

Example 13. Assume that $X \sim B(n, p)$. Complete each statement.

(1) $P(x \geq 8) = 1 - P(x \leq \quad)$
(2) $P(x > 6) = 1 - P(x \leq \quad)$
(3) $P(x < 7) = P(x \leq \quad)$

Note: These equations are valid only for *discrete* random variables.

Assume $X \sim B(n, p)$. Use the calculator keys $< 2^{nd} >$ <VARS> to access functions involving probability distributions ("DISTR"). The calculator can find the probability that x IS a particular value (binom**p**df), or that x is AT MOST a particular value (binom**c**df):

$$P(x = 5) = \text{A:binompdf}(n, p, 5),$$
$$P(x \leq 5) = \text{B:binomcdf}(n, p, 5).$$

Example 14. FINDING BINOMIAL PROBABILITIES. According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

(1) What is the probability distribution of X?
(2) Use your calculator to find the following probabilities:
    (a) the probability that 25 adults in the sample prefer saving over spending
    (b) the probability that at most 20 adults prefer saving
    (c) the probability that more than 30 adults prefer saving

TryIt 4.14.

Example 15. FINDING BINOMIAL PROBABILITIES. The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

(1) What is the probability distribution of X?
(2) Using the formulas and appropriate notation, calculate the mean and standard deviation of X.
(3) Interpret the mean you found in the context of this problem.
(4) Find the probability that at most five people develop pancreatic cancer.
(5) Find the probability that eight or more people develop pancreatic cancer.
(6) Is it more likely that 5 or 6 people will develop pancreatic cancer? Justify your answer numerically.

Example 4.15.

The formula for calculating binomial probabilities "by hand" isn't very difficult. Assume $X \sim B(n, p)$. Then

$$P(x = k) = \binom{n}{k} p^k q^{n-k},$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Depending on how much probability we did in Chapter 3, it may or may not make sense to show you why this formula works.

3.4. **The Shape of the Binomial Distribution.**

Example 16. A September 2015 article on CNET.com claims that 40% of mothers avoid family photos because they don't like how they look. Suppose this number is true and that you survey 10 mothers with this question. Let X be the number of moms out of 10 who say they avoid family photos for this reason. Then $X \sim B(10, 0.4)$.

(1) The following are a simple random sample of size 50 from this distribution. In other words, each of these numbers represent the number of moms who say they avoid family photos from a random sample of 10. (So it's as if we have conducted 50 samples of size 10.)

6	4	3	2	6	5	3	6	4	6
4	3	3	5	4	2	5	7	2	4
3	3	3	4	3	5	3	5	5	4
3	5	4	5	3	4	5	5	4	1
3	3	3	6	5	5	4	3	1	5

   (a) Draw a histogram of the sample data with the following class boundaries: $-0.5, 0.5, 1.5, 2.5, \ldots, 10.5$.
   (b) Describe the distribution of the sample data based on the histogram. (What's the shape of the histogram? Which is the most frequent data value? Is the data fairly symmetric or skewed?)
   (c) Find the mean and sample standard deviation of the sample data.
(2) Now focus on the theoretical distribution of $X \sim B(10, 0.4)$.
   (a) Complete the probability distribution table for X. (Use "2^{nd}, VARS, binompdf(10, 0.4), STO>, 2^{nd}, 2" to load the probabilities into L2.

You can manually enter the x values into L1, if you wish.)

x	P(x)
0	
1	

(b) Draw a histogram using class boundaries $-0.5, 0.5, 1.5, 2.5, \ldots, 10.5$ and bar heights corresponding to the probability distribution of X.

(c) Describe the distribution of random variable $X \sim B(10, 0.6)$ based on this histogram. (What's the shape of the histogram? Which is the most frequent data value? Is the data fairly symmetric or skewed?)

(d) Find the mean and standard deviation of the probability distribution using the formulas.

(3) Compare the distribution of the sample data from (1) to the theoretical binomial probability distribution from (2) by comparing: the shapes of the histograms, the most common data values, the means, and the standard deviations.

(a) Are the distributions exactly the same? Why or why not?

(b) Are the distributions similar? Why or why not?

Summarizing Example 16: The probability distribution of a random variable provides the characteristics (shape of histogram, mean, standard deviation) of the "perfect" data set following that distribution.

Actual random samples of data inevitably vary from this "perfect" distribution. According to The Law of Large Numbers, the larger the sample size, the better the sample data will fit the probability distribution.

CHAPTER 5: CONTINUOUS PROBABILITY DISTRIBUTIONS

LECTURE NOTES FOR INTRODUCTORY STATISTICS[1]

Daphne Skipper, Augusta University (2016)

1. CONTINUOUS PROBABILITY FUNCTIONS (PROBABILITY DENSITY FUNCTIONS)

In the last chapter, we discussed discrete random variables and their probability distribution functions. We were able to define these functions using tables because we were able to list out all of the possible data values. However, we can't list all the possible outcomes of a continuous random variable, because the range of data values are all the values in an interval on the number line.

The probability distribution of a continuous random variable, X, is defined by its **probability density function (pdf)** or **density curve**, $f(x)$. The most important fact about density curves is:

Area under the density curve, $f(x)$, corresponds to probability.

Example 1. PROBABILITY = AREA UNDER DENSITY CURVE. Draw the "perfect bell-shaped distribution" using a smooth line (rather than histogram bars) and with the mean, $\mu = 4$ in the center of the curve on the $x-$axis. This is what the density curve for the normal distribution looks like.

(1) Shade the area that equals $P(3 < x < 5)$
(2) Shade the area that equals $P(x < 2.5)$
(3) Shade the area that equals $P(x > 6)$

NOTE: For *continuous* distributions, $P(3 < x < 5) = P(3 \leq x \leq 5)$. In fact, $P(x = 3) = 0$ because the area under the curve at a single number has no width. ALSO NOTE: Calculus is required to find areas under the normal density curve. We will use technology (calculators) to find these areas.

A function $f(x)$ must have the following **properties** to be considered a **"valid" probability density curve**:

(1) the entire curve must be above the $x-$axis,
(2) the TOTAL area under the curve must be _____.

A few more facts about density curves:

- The $x-$axis represents the range of data values.
- The $y-$axis has decimal values, like relative frequencies.
- The curve is smooth, but otherwise matches a "perfect histogram" of the data.
- We will find probabilities of *intervals* of numbers, rather than exact numbers:
  - $P(3 < x < 5)$ = the probability that a randomly selected data value is between 3 and 5

[1]These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

- $P(x < 2.5)$ = the probability that a randomly selected data value is at most 2.5
- $P(x > 6)$ = the probability that a randomly selected data value is more than 6

2. The Uniform Distribution

The uniform distibution is the simplest continuous probability distribution.

X is a **uniform random variable** if the possible data values range over an interval of the number line, and every number in that interval is *equally likely* (not just whole numbers, but decimals too).

Notation: $X \sim U(a, b)$ means that random variable X has a uniform distribution over the interval $[a, b]$.

Example 2. Recognizing a uniform distribution. The data in the following table are 55 smiling times, in seconds, of eight-week-old babies.

10.4	19.9	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	23.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	23.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

Complete the frequency/relative frequency table for the smiling times data.

x	frequency	rel freq
0-4	6	0.109
4-8	11	0.2
8-12	13	0.236
12-16	8	
16-20	10	
20-24		

(If a data value falls on a boundary value, put it into the *lower* class.)

(1) Use words to describe random variable X: $X =$
(2) Is X continuous or discrete?
(3) Does X appear to have a uniform distribution? In other words:
  (a) What appears to be the range of possible data values, x?
  (b) Do all of the data values in this range appear to be equally likely?
(4) $X \sim$

Since all the data values are equally likely, the **mean** of $X \sim U(a, b)$ is simply the midpoint of the range of possible data values:

$$\mu = \frac{a+b}{2}.$$

The **standard deviation** of $X \sim U(a, b)$ less intuitive, but still a simple calculation:

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Example 3. Mean of the uniform distribution. Suppose $X \sim U(0, 24)$, the length of smiles (in seconds) of eight-week-old babies (from Example 2).

(1) Find the theoretical mean smiling time of this eight-week-old baby. Use the correct notation.
(2) Use the frequency table from Example 2 to estimate the mean of the sample data. Use the correct notation.
(3) Are the theoretical and sample means the same? Why or why not?

Example 5.2 continued.

Example 4. HEIGHT OF A UNIFORM DENSITY CURVE. Suppose $X \sim U(0, 24)$, the length of smiles (in seconds) of eight-week-old babies (from Example 2).

(1) Recall that a density "curve" is a smooth line that follows the shape of a "perfect histogram". Sketch the density curve for X. Make sure the curve starts and stops at the correct x values, but don't worry about the height of the curve for now.
(2) The area under the curve must be _____.
(3) The area under the curve has the shape of a _____. Use what you know about this shape to find the height of the curve and label it on the $y-$axis.
(4) Use what you learned from this example to make a general formula for the height of the probability density curve for uniform random variable $X \sim U(a, b)$.

Example 5.2 continued.

If $X \sim U(a, b)$, the probability density curve $f(x)$ is a _____ from ___ to ___ on the $x-$axis at a height of _____. In mathematical notation, we write

$$f(x) = \frac{1}{b-a}, \text{ where } a \leq x \leq b.$$

(You will use this notation in the homework.)

For continuous distributions, probabilities are areas under the density curve. In particular, **uniform distribution *probabilities*** are ***areas of rectangles***, which we know all about.

Example 5. UNIFORM DISTRIBUTION PROBABILITIES. Suppose $X \sim U(0, 24)$, the length of smiles (in seconds) of eight-week-old babies (from Example 2). As you work through the following, use the correct notation for probabilities.

(1) What is the probability that a randomly chosen eight-week-old baby smiles between 2 and 18 seconds?
(2) Find the 90^{th} percentile for an eight-week-old baby's smiling time.
(3) Find the probability that an eight-week-old baby smiles more than 12 seconds knowing that the baby smiles more than eight seconds. *Hint: Consider another uniform distribution over the restricted sample space of 8 to 24 seconds.*

Example 5.3.

Example 6. Suppose the time it takes a nine-year-old child to eat a donut is uniformly distributed between 0.5 and 4 minutes, inclusive. Let $X =$ the time, in minutes, it takes a nine-year-old child to eat a donut: $X \sim$ _____. Use the correct notation for probabilities.

(1) Find the mean amount of time it takes a nine-year-old child to eat a donut. Use the correct notation for the mean.
(2) The probability that a randomly selected nine-year-old child eats a donut in at least two minutes is _____.
(3) Find the probability that a different nine-year-old child eats a donut in more than two minutes, given that the child has already been eating the donut for more than 1.5 minutes.

Example 5.5

CHAPTER 6: THE NORMAL DISTRIBUTION

LECTURE NOTES FOR INTRODUCTORY STATISTICS[1]

Daphne Skipper, Augusta University (2016)

Data following a **normal distribution** have histograms that are approximately symmetrical and bell-shaped. Most data values are clumped close to the mean. As data values get farther from the mean, they are less and less common.

A normally distributed random variable is completely described by its mean and standard deviation. We use the notation $X \sim N(\mu, \sigma)$ to say that random variable X follows a normal distribution with mean μ and standard deviation σ. Examples of data that follow a normal distribution:

- men's heights
- women's shoe sizes
- women's pulse rates

Example 1. RECOGNIZING THE NORMAL DISTRIBUTION. A sample of **n = 1000** women's pulse rates begins with the following data values:

74 87 60 77 67 68 69 73 87 81 ...

The full data set has the following relative frequency table:

Pulse	Rel Freq
47-54	.01
54-61	.06
61-68	.13
68-75	.20
75-82	.24
82-89	.19
89-96	.10
96-103	.05
103-110	.02

(1) Draw a relative frequency histogram for this data set.
(2) Does this data appear to follow a normal distribution?
(3) What would you guess is the mean of the sample data?
(4) The sample mean and sample standard deviation are: $\overline{x} = 77.7$ and $s = 11.8$. Mark these values on the x-axis of the histogram.
(5) Suppose random variable X = woman's pulse rate.
   (a) We could estimate that $X \sim$ ___(___, ___).
   (b) Sketch the approximate density curve for X over the histogram.
(6) Give an example of a pulse rate that would be considered unusually high.
(7) Give an example of a pulse rate that would be considered unusually low.

Data from "Elementary Statistics" by M. Triola.

A common error that researchers make is to assume a sample arises from a normal distribution when in fact it does not. Because the sample in the last example is

[1]These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

fairly large, its histogram has a very clear bell-shape. For smaller data sets, the bell-shape may not be as clearly apparent. There are methods to help us decide if it is "safe" to assume that data arise from a normally distributed population, such as a *normal quantile plot*, but ultimately this is a gray area and relies on the discretion of the researcher.

1. The Standard Normal Distribution

The ***standard* normal distribution** is a normal distribution with mean $\mu = \mathbf{0}$ and standard deviation $\sigma = \mathbf{1}$: $Z \sim N(0, 1)$. We use Z instead of X for a standard normal random variable because standard normal data values are **z scores**.

This is a good time to review a few **properties of z scores**:

- the z score of a data value tells us how many standard deviations from the mean that data value is,
- z is positive if the data value x is above the mean,
- z is negative if x is below the mean, and
- z is 0 if x is the mean.
- z scores *standardize* data values, allowing us to compare data values in different normal distributions to see which is more common or extreme relative to its own mean and standard deviation.

Suppose $X \sim N(\mu, \sigma)$. If we need to **find the z score** associated with the data value x, we use the formula

$$z = \frac{x - \mu}{\sigma}.$$

If we need to **find the data value** that has a particular z score, we use the formula

$$x = \mu + z\sigma.$$

Example 2. Comparing data values using z scores. SAT scores and ACT scores are both normally distributed. SAT scores have a mean of 1026 and a standard deviation of 209. ACT scores have a mean of 20.8 and a standard deviation of 4.8. A student takes both tests and scores 1130 on the SAT and 25 on the ACT.

(1) Suppose X = score on the SAT. Then $X \sim$ ___(___, ___).
(2) Suppose Y = score on the ACT. Then $Y \sim$ ___(___, ___).
(3) Compare the test scores using z scores.

algebra.com

If X is a random variable and has a normal distribution with mean and standard deviation , then the **Empirical Rule** says the following:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean).

The empirical rule is also known as the **68-95-99.7 rule**.

Section 6.1.

Example 3. The 68-95-99.7 Rule. From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

(1) About 68% of the y values lie between what two values?
(2) About 95% of the y values lie between what two values?
(3) About 99.7% of the y values lie between what two values?

Example 6.6.

2. Using the Normal Distribution

Recall that for continuous random variables, probabilities are the same as area under the density curve. The calulator (or various computer programs) helps us to calculate these areas.

STRATEGY: For each question, take a minute to figure out what is given and what is unknown. Then sketch a normal diagram, shade the relevant area, and mark what is given and what is unknown.

- Finding a probabilty: cutoff data value(s) x given; area unknown.
- Finding a percentile/quartile: area given; cutoff data value x unknown.
- Given a percentage: area given; cutoff data value(s) unknown.

CALCULATOR FUNCTIONS: Suppose $X \sim N(\mu, \sigma)$.

- If the AREA is UNKNOWN, use 2^{nd}, VARS, **normalcdf** to find the area.
  - normcdf$(a, b, \mu, \sigma) = P(a \leq x \leq b) = P(a < x < b)$
  - normcdf$(-10^9, b, \mu, \sigma) = P(x \leq b) = P(x < b)$
  - normcdf$(a, 10^9, \mu, \sigma) = P(x \geq a) = P(x > a)$
- If the AREA is KNOWN, use 2^{nd}, VARS, **invnorm** to find the cutoff data value that is the (right) boundary value of the given area (which is unbounded on the left).
  - invnorm(area cumulative from left, μ, σ) = cutoff data value (bounding the right side of the area)

  Important: The area entered into **invnorm** is assumed to be cumulative from the left. It is important to draw a picture and add/subtract to figure out the area to the left of the unknown data value before proceeding.

Example 4. Normal Distribution Probabilities. A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour. For each of the following, sketch the graph and solve.

(1) Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
(2) Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Example 6.9.

Example 5. Normal Distribution Probabilities. There are approximately one billion smartphone users in the world today. In the United States the ages

13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively. For each of the following, sketch the graph and solve.

(1) Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.
(2) Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.
(3) Find the 80th percentile of this distribution, and interpret it in a complete sentence.
(4) Forty percent of smart phone users are over the age of _____.

Example 6.10.

Example 6. Normal Distribution Probabilities. A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm. For each of the following, sketch the graph and solve.

(1) Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm.
(2) The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.
(3) Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Example 6.12.

CHAPTER 7: THE CENTRAL LIMIT THEOREM

Lecture Notes for Introductory Statistics[1]

Daphne Skipper, Augusta University (2016)

1. The Central Limit Theorem For Means

1.1. **CLT Setup.** Suppose we know that in 2005, the average number of points scored in an NFL game (combined score of both teams) was 42 with a standard deviation of 8 points. We wonder if the scoring has changed over the last decade, so we take a random sample of thirty games in the 2015 season and calculate a sample mean of 46.8 points per game.

Of course this makes us suspect that the average number of points per game has increased, but we only have partial 2015 data, so it is not enough simply to compare the means. We rephrase the question:

"Is there *statistical evidence* that there are more points scored per game now than there were in 2005?"

To answer this question, we have to decide if the sample mean of 46.8 is very different from what we would expect if the (population) mean number of points is still 42, like in 2005. In other words, we need to see if $P(\overline{x} \geq 45)$ is very small, assuming the population mean number of points per game is still 42. But to calculate this probability, we need the distribution of $\overline{X}$, rather than the distribution of X! This is where the Central Limit Theorem comes in.

In this example, we have the random variables X and $\overline{X}$:

X = the number of points scored in an NFL game in 2015,
$\overline{X}$ = the mean (average) number of points scored in a random sample of $n = 30$ NFL games in 2015.

The Central Limit Theorem says that if X has mean $\mu = 42$ points (our assumed average) and standard deviation $\sigma = 8$ points (our assumed standard deviation), then $\overline{X} \sim N\left(42, \frac{8}{\sqrt{30}}\right)$.

1.2. **CLT Statement.** Under certain conditions, the Central Limit Theorem provides the distribution of the random variable defined to be the means of samples of a certain size from a distribution whose mean and standard deviation are known.

The **Central Limit Theorem** says that if

X = a random variable (any distribution) with mean μ_X and standard deviation σ_X, and
$\overline{X}$ = the mean of a random sample of n data values from X,

[1]These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

then the distribution of $\overline{X}$ approaches the distribution $N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$ as n gets large.

Let's take a closer look at the requirements for the Central Limit Theorem. The following are true:

- Regardless of n, it is always true that $\mu_{\overline{X}} = \mu_X$ and $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$.
- Regardless n, if X is normally distributed, then $\overline{X}$ is also normally distributed.
- If X is not normally distributed, the distribution of $\overline{X}$ *approaches* a normal distribution as n gets large.
- How big is "large"? Many textbooks recommend that if X is normally distributed OR $\mathbf{n \geq 30}$, then we can assume $\overline{X}$ is normally distributed .

1.3. **CLT Intuition.** The Central Limit Theorem is actually quite intuitive. For larger and larger sample sizes, we would expect (by the Law of Large Numbers) for the sample mean $\overline{x}$ to get closer and closer to the population mean μ_X. So it makes sense that the mean of $\overline{X}$ is the same as the mean of X, and the bigger the sample size, the smaller the standard deviation (or standard error) of $\overline{X}$. In other words, the sample mean is a better and better approximator of the population mean as the sample size increases.

2. The Central Limit Theorem For Sums

We aren't doing this section.

3. Applying the Central Limit Theorem (For Means)

In this section we apply the Central Limit Theorem the find probabilities and percentiles involving $\overline{X}$.

Example 1. Finding probabilities involving X and $\overline{X}$. The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of n = 60 is drawn randomly from the population.

(1) $X =$
(2) $X \sim$
(3) $\overline{X} =$
(4) $\overline{X} \sim$
(5) Find the probability that a single student takes between 2 and 3 hours.
(6) Find the probability that the sample mean of 60 students is between 2 and 3 hours.

TryIt 7.2.

Example 2. Probabilities and Percentiles involving $\overline{x}$. In a study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users was 34 years. Suppose the standard deviation is 15 years. Take a sample of size n = 100.

(1) What are the mean and standard deviation for the sample mean ages of tablet users?
(2) What does the distribution look like?
(3) Find the probability that the sample mean age is 30 or older.

(4) Find the 10th percentile for the sample mean age.
(5) Find the 95th percentile for the sample mean age.

Example 7.3.

Let's return to the example from the beginning of the notes for this chapter...

Example 3. STATISTICALLY SIGNIFICANT RESULT? Suppose we know that in 2005 the average number of points scored in an NFL game (combined score of both teams) was 42 with a standard deviation of 8 points. We wonder if the scoring has changed over the last decade, so we take a random sample of 30 games in the 2015 season and calculate a sample mean of 46.8 points per game. Let $X =$ the number of points scored in a 2015 NFL game.

(1) We will assume that X has a mean and standard deviation equal to those from the 2005 NFL season. (Why are we making this assumption about X?)
(2) Based on our assumption about X, $\overline{X} \sim$
(3) Based on our assumption about X, $P(\overline{X} \geq 46.8) =$
(4) Describe the significance of the probability from part (2).
(5) Is there statistical evidence that there are more points scored now than in 2005? Explain.

4. STATS LAB

This is a good time for a project to see the Central Limit Theorem in action. The Cookie Recipe Stats Lab at the end of Chapter 7 is a good option.

CHAPTER 8: CONFIDENCE INTERVALS

Lecture Notes for Introductory Statistics[1]

Daphne Skipper, Augusta University (2016)

"Confidence Intervals" is our first topic in **inferential statistics**. In this chapter, we use sample data to estimate an unknown population parameter: either population mean (μ) or population proportion (p). (See why it's called *inferential* statitistics?)

Population parameters are generally unknown, because it is hard to have population data. We have talked about how we can use $\bar{x}$ to estimate μ. In fact, we say that $\bar{x}$ is the **point estimate** for μ; $\bar{x}$ is our single best guess at μ.

In this Chapter, we estimate population parameters by providing a **confidence interval** of numbers that we are, say, 95% certain captures the value of the population parameter. The interval is centered at the point estimate and extends in both directions by the **margin of error**:

$$\text{confidence interval for parameter} = \text{point estimate} \pm \text{margin of error}.$$

In this scenario, "95%" is the **confidence level, CL**. The margin of error depends on the confidence level. (If we want to be more confident that the interval contains the population parameter, what must we do to the margin of error?)

1. A Single Population Mean (Confidence Interval) using the Normal Distribution

For the case of estimating population mean μ, the point estimate is $\bar{x}$, and the margin of error is known as the **EBM**, or *error bound for population mean*. Thus, the confidence interval for population mean has the form

$$\text{confidence interval for } \mu = (\bar{x} - EBM, \bar{x} + EBM)$$

Notice that the two equations above demonstrate two different, but equally valid, ways to express confidence intervals: using $\pm$ notation or interval notation.

As noted above, margin of error (EBM) depends on the confidence level (CL). We choose the EBM so that we can say something like: "we are 95% certain that this confidence interval contains the true population parameter." Actually, a more accurate statement would be "95% of samples of this size from the population produce confidence intervals that contain the true value of the population parameter." (Remember that we use samples from the population to do our estimating.)

Conversely, we could say that "5% of samples of this size from the population produce confidence intervals that DO NOT contain the true value of the population parameter." This value, 5% is known as α ("alpha"). So,

$$\alpha = 1 - \text{CL}$$

is the probability that our confidence interval DOES NOT contain the true value of the population parameter. Note that CL and α are both *probabilities* and may be expressed as decimals or percents.

[1]These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

A useful interpretation of a confidence interval estimate is to determine if there is **statistically significant evidence** that a population parameter is above or below some value. For example, suppose (34.5%, 36.2%) is a 95% confidence interval estimate for the percentage of kindergarteners who bring a lunch box to school. There is statistically significant evidence (at the 95% confidence level) that the percentage of kindergarteners who bring a lunch box to school is greater than 34%, because according to the confidence interval all of the possible values of p are over 34%. However, there is NOT statistically significant evidence (at the 95% confidence level) that more than 35% of kindergarteners bring a lunch box to school, because there are possible values of p in the confidence interval that are below 35%.

Example 1. Confidence Interval Form. Suppose we have collected data from a sample. We know the sample mean is seven, and the error bound for the mean is 2.5 for a confidence level of 95%.

(1) $\bar{x} =$
(2) $EBM =$
(3) Then confidence interval is _____ ±_____ or (_____, _____)
(4) We estimate with _____ confidence that the true value of the population mean μ is between _____ and _____.
(5) Is there statistically significant evidence at the 95% conficence level that the population mean is less than 8?

Example 8.1.

We turn our focus to the margin of error, EBM; where does this number come from? Draw a normal distribution with μ at the center. Suppose we want a 90% confidence interval. Shade an area of 90% in the center. (What is α?_____ Where does α show up in the drawing?_____)

This normal distribution is NOT X, but is $\bar{X}$ (which we can often assume is normally distributed due to the Central Limit Theorem), because we want the data values to be $\bar{x}$'s, sample means. In our drawing, we can see that 90% of samples (of the appropriate size) will have means that land on the numberline *inside* the shaded area, and 10% of samples will have means that lie on the numberline *outside* the shaded area (in the "tails").

We need for the samples whose means fall within the shaded area to result in confidence intervals that capture the true value of the population mean, which is right in the center. So on the drawing, how wide must the EBM be?

More precisely, to capture the central 90% of data, we must go out 1.645 standard deviations from the mean on either side. So, for this example, $EBM = 1.645 * (\text{standard deviation})$. Due to the Central Limit Theorem, we know that the standard deviation of $\bar{X}$ is _____. So we have $EBM = 1.645 * \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population of interest and n is the size of the sample we have collected.

(Yes, it IS kind of reckless to assume we know σ when we don't know μ, but we are going to assume we have that information for now. We will handle the dilemma of unknown σ in the next section.)

Example 2. Find the percentile ($P_?$) of the z-score and the z-score for calculating the margin or error (EBM) for each confidence level. Remember that z-scores come from a *standard* normal distribution: $Z \sim N(0, 1)$.

(1) 90%: z-score = $P_\square =$

(2) 95%: z-score = $P_{\square} =$

(3) 99%: z-score = $P_{\square} =$

In general,

$$\text{z score} = P_{(CL+\frac{\alpha}{2})} = \text{invnorm}(CL + \frac{\alpha}{2}, 0, 1)$$

where $\alpha = 1 - CL$.

To summarize, if we want to use a random sample of size n to generate a CL confidence interval estimate for the population mean μ of a population that has standard deviation σ,

$$\text{confidence interval estimate} = \bar{x} \pm EBM = (\bar{x} - EBM, \bar{x} + EBM)$$

where

$$EBM = \text{z-score} * \frac{\sigma}{\sqrt{n}}.$$

Example 3. Calculating a Confidence Interval Estimate for μ. Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and has a sample mean (average score) of 68. Follow the steps to find a 90% confidence interval estimate for the population mean exam score (the average score on all exams).

(1) Identify $\bar{x}$, σ, CL, and α.
(2) Find the appropriate z-score.
(3) Calculate the margin of error, EBM.
(4) Write the confidence interval in the $\pm$ form.
(5) Write the confidence interval in the interval notation form.
(6) Check your work by finding the confidence interval using the **calculator function STAT, TESTS, 7:ZInterval**.
(7) Plot the confidence interval on a number line. Label the sample mean and the EBM.
(8) Interpret the confidence interval in the *context of this problem.*
(9) Is there statistically significant evidence, at the 90% confidence level, that the population mean score is below 70?

Please do the following to be turned in at the next class meeting. Do all calculations by hand, showing your work, checking your work with the calculator function "ZInterval."

(1) Find the z-score that would be required to calculate the margin of error (EBM) for a 93% confidence level.

(2) Explore changing the confidence level:
   (a) Find the EBM and confidence interval for Example 3, but with a confidence level of 95% (keeping everything else the same).

   (b) Find the EBM and confidence interval for Example 3, but with a confidence level of 99%.

   (c) How does increasing the confidence level affect the margin of error, EBM? The width of the confidence interval?

   (d) Explain in your own words why the margin of error is affected in this way when the confidence level is increased.

(3) Explore changing the sample size:
   (a) Find the EBM and confidence interval for Example 3, but with a sample size of 100.

   (b) Find the EBM and confidence interval for Example 3, but with a sample size of 500.

   (c) How does increasing the sample size affect the margin of error, EBM? The width of the confidence interval?

   (d) Explain in your own words why the margin of error is affected in this way when the sample size is increased.

2. A Single Population Mean (Confidence Interval) using the Student t Distribution

In the last section, we used a random sample from a population to create a confidence interval for the population mean μ of a population. The confidence interval had the form

$$\bar{x} \pm EBM = \bar{x} \pm (\text{z-score})\frac{\sigma}{\sqrt{n}}.$$

In this section we are still estimating μ with a confidence interval. The difference is that in this section *we DO NOT assume the standard deviation σ of the population is known.* (Is more likely that σ is known or unknown?)

How does NOT KNOWING σ affect the construction of the confidence interval? Well certainly it does not affect the first part of the formula: $\bar{x}$. However, unknown σ impacts the formula for the EBM in two ways:

(1) Just like $\bar{x}$ is the best point estimate for μ, the *sample* standard deviation s is the best point estimate for the *population* standard deviation σ. So the best we can do in the formula is to put s in the place of σ.
(2) As you might guess, using s in the place of σ introduces some error into the EBM, but we don't know if it errs by making the EBM too large or too small. If $s < \sigma$, a 95% confidence interval will be too small to capture μ 95% of the time. Statisticians counterbalance this possible error by using a t-score instead of a z-score, which makes the EBM just a tad larger.

When creating a **confidence interval estimate for population mean** μ, if the population standard deviation σ **is UNKNOWN**,

$$EBM = (t\text{-score})\frac{s}{\sqrt{n}}.$$

Obviously we now need to know more about what a t-score is. As stated above, the t-score for a CL is smidge larger than a z-score for the same CL, in order to offset any error introduced by using s in place of σ. How much larger? That depends on the sample size:

- If the sample size is fairly large, s is expected to be very close to σ, the error introduced by using s will be small, and the t-score will be only slightly larger than the z-score.
- On the other hand, if the sample size is small, the error introduced by using s to estimate σ will be bigger, so the t-score has to be larger relative to the z-score to off-set that error.

Just like z-scores arise from a Z distribution, t-scores arise from a T distribution. Here are a few important facts about the T distribution:

- The T density curve is bell-shaped and has mean 0 and standard deviation 1, just like the standard normal distribution, Z.
- The precise shape of the T density curve depends on the sample size, or more precisely, on the **degrees of freedom, df**:

  $$\text{df} = n - 1.$$

  The higher the degrees of freedom, the more closely the T distribution matches the standard normal distribution, Z.
- "William S. Goset (1876−1937) of the Guinness brewery in Dublin, Ireland ran into [the problem of error introduced by small samples]. His experiments with hops and barley produced very few samples...This problem led him to "discover" what is called the Student's t-distribution. The name

comes from the fact that Gosset wrote under the pen name "Student"."
Chapter 8 excerpt.

How do we find a t-score? There are two ways, depending on which calculator you have.

(1) **If you have a TI-84+**, your calculator has a function **invT**, which works just like the "invnorm" function, except that you have to enter the degrees of freedom, df:

$$t\text{-score} = \text{invT(area to the left, } df\text{)}, \text{ where } df = n - 1.$$

(2) **If you have a TI-83+**, you have to use a **table of t-scores**, which you can find in the book or on the internet. If using the table from the book, scan across the top to find the "area to the left" of your t-score. Scan down to find the degrees of freedom, $df = n - 1$.

NOTE: You will often see a subscript on t and z. The subscript is the "tail probability" associated with the score, which is $\alpha/2$. For example, a confidence level of 95% is associated with the z-score $z_{0.025} = \text{invnorm}(0.975, 0, 1)$.

Example 4. CALCULATING t-SCORES. Find the t-score associated with each confidence level and sample size combination. Use invT(area, df) if you have a TI-84+ and the table otherwise.

(1) CL = 90%, $n = 6$
(2) CL = 90%, $n = 100$
(3) CL = 95%, $n = 6$
(4) CL = 95%, $n = 100$

Summary: Confidence Interval for Population Mean μ
Requirements: The margin of error (EBM) calculation requires that $\bar{X}$ is normally distributed. By the Central Limit Theorem, we require that $n > 30$ or X is normally distributed, which we verify by checking to see if the sample appears to be *approximately* normally distributed.

- If σ is KNOWN: $\bar{x} \pm (z\text{-score})\frac{\sigma}{\sqrt{n}}$
- If σ is UNKNOWN: $\bar{x} \pm (t\text{-score})\frac{s}{\sqrt{n}}$

These are the **calculator functions** that calculate confidence intervals for μ.

- If σ is KNOWN: STAT, TESTS, ZInterval
- If σ is UNKNOWN: STAT, TESTS, TInterval

What you should know about using these calculator functions:

(1) The calculator prompts you on what to enter, but you have to know what the symbols mean.
(2) You have a choice of "Data" or "Stats". Choose the "Data" option if you have the data stored in a list on your calculator. (If you have the data in a single list (not as a frequency table in two lists), enter "Freq: 1".)
(3) The confidence interval you get back is in *interval notation*. If you need to recover $\bar{x}$ and EBM from the confidence interval, remember that $\bar{x}$ is the midpoint of the interval and EBM is half the width of the interval. For a confidence interval (a, b),

$$\begin{aligned} \bar{x} &= \tfrac{a+b}{2} \\ EBM &= \tfrac{b-a}{2} \end{aligned}$$

Example 5. Confidence Interval for μ (σ unknown). You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data. 8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10. Calculate the confidence interval using the formula, then check it with the appropriate calculator function.
TryIt 8.8.

Example 6. Confidence Interval for μ. A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

(1) Use appropriate symbols to label all of the information, including X and $\bar{X}$ and their distributions.
(2) Which distribution should you use: T or Z?
(3) Use a calculator function to make a 95% confidence interval for the mean weight of all heads of lettuce.
   (a) interval notation:
   (b) $\pm$ notation:
(4) Do we have evidence, at the 95% confidence level, that the (population) mean weight of a head of lettuce is under 2.3 pounds? More than 2.15 pounds?

Ch 8 Exercises.

3. A Population Proportion (Confidence Interval)

In this section we switch from estimating μ to estimating p, **population proportion**.

When estimating population mean μ, the data is quantitative and often continuous. The units on the confidence interval bounds match the units of the data. For example: "With 95% confidence, a person using hypnotherapy sleeps is between 8.17 and 8.46 hours per night on average."

When **estimating population proportion** p, the **data is categorical (with two categories of interest)**. The units on the confidence interval bounds are percentages. For example: "With 95% confidence, between **8.7%** and **9.5%** of kindergarteners know how to read when they start school."

The scenario where the data is categorical with two categories should sound familiar, like a binomial random variable. Recall that if $X \sim B(n, p)$, then

$$\begin{aligned} X = &\quad \text{the number of successes out on } n \text{ trials,} \\ \mu_X = &\quad np, \\ \sigma_X = &\quad \sqrt{npq}, \text{ and} \end{aligned}$$

the probability distribution of X is bell-shaped, like the normal distribution.

However, what we are actually interested in here is not the *number* of successes, x, but the *proportion* of successes $p' = \frac{x}{n}$. Now, if you divide all the x data values by n, you have new random variable $\frac{X}{n}$ = the *proportion* of success in n trials. In other words, $\frac{\mathbf{X}}{\mathbf{n}}$ represents the **population of sample proportions, p′** (much like $\bar{X}$ represents the population of sample means).

NOTATION:

$$\begin{aligned} p = &\quad \textbf{population} \text{ proportion} \\ p' = &\quad \textbf{sample} \text{ proportion} \end{aligned}$$

The random variable $\frac{X}{n}$ has mean $\frac{np}{n} = p$ and standard deviation $\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$. The histogram of $\frac{X}{n}$ matches that of X: it is bell-shaped. In fact, the histogram has such a nice bell-shape that as long as the sample size n is fairly large and p is not very close to 0 or 1, we can approximate the population of sample proportions by a normal distribution:

$$\frac{X}{n} \sim N\left(p, \sqrt{\frac{pq}{n}}\right).$$

Recall our general formula for confidence intervals:

$$\text{point estimate} \pm \text{margin of error.}$$

In this scenario, the best point estimate for population proportion p is sample proportion p'. The margin of error is called the **error bound for the proportion, EBP** and follows the same general formula as the EBM:

$$\text{EBP} = (\text{z-score})(\text{standard deviation}) = (\text{z-score})\left(\sqrt{\tfrac{pq}{n}}\right)$$

Notice that the formula requires unknown p. The best we can do is to replace p by it's best point estimate, p', and q by $q' = 1 - p'$. We calculate EBP as follows:

$$\text{EBP} = \quad (\text{z-score})\left(\sqrt{\tfrac{p'q'}{n}}\right).$$

Now, our **confidence interval for population proportion p** is

$$p' \pm (\text{z-score})\left(\sqrt{\frac{p'q'}{n}}\right),$$

where

$$p' = \frac{x}{n} = \frac{\text{number of successes}}{\text{sample size}}$$

and

$$q' = 1 - p'.$$

The **calculator function** that finds confidence intervals for population proportion directly is:

$$\text{STAT, TESTS, 1-PropZInt.}$$

It is very simple to use, as long as you know what the symbols mean. (Can you figure out what all the parts of the name "1-PropZInt" mean?)

Example 7. CONFIDENCE INTERVAL FOR POPULATION PROPORTION. Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets. Use the formula and check your work using the calculator function 1-PropZInt. Is there statistically significant evidence at the 95% confidence level that less than 43% of all people own tablets? TryIt 8.10.

Interpreting Confidence Interval Overlap. We say that there is statistical evidence that the means (or proportions) are different as long as neither confidence interval captures the point estimate of the other confidence interval.

Example 8. Interpreting confidence interval "overlap." Suppose we have confidence intervals for the population mean scores on a statistics test by gender. What can we conclude about the population mean scores of boys and girls in each case: is there statistical evidence that the mean scores are different?

(1) Boys: 73 ± 2, Girls: 78 ± 2
(2) Boys: 73 ± 3, Girls: 71 ± 2
(3) Boys: 73 ± 3, Girls: 69 ± 2

Sample Size Determination. An important question when conducting statistical analysis is, **"how large must the sample be?"** Often getting larger samples is expensive, labor intensive, and/or time consuming. Fortunately, it is possible to analyze how large your sample must be to achieve the desired level of confidence and margin of error.

The strategy for determining sample size is to use the margin or error formula, EBM and EBP, and solve for n.

One thing to recognize is that increasing sample size decreases margin of error. We will want "at most" some amount of error. If our sample size falls short, we will have too much error. Therefore, with sample size, we always err on the side of too many samples. This reasoning accounts for the following two conventions:

- Set $p' = q' = 0.5$ in the formula. (Naturally, sample size calculation occurs before data collection, so we don't yet have p'. The choice of $p' = 0.5$ maximizes $(p')(q')$, preventing an underestimate of the required sample size.)
- Round UP sample size calculations to the next whole number. (Sample size is a whole number. If we round down, we underestimate the required sample size.)

Example 9. Required Sample Size for a Confidence Interval for p. Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones. Example 8.14. (Barbara Illowsky & Susan Dean. Introductory Statistics. OpenStax College, 2013.)

(1) $EBP = (\text{z-score})\left(\sqrt{\frac{p'q'}{n}}\right)$. Our strategy is to fill in everything except n, then to solve for n.
   (a) Find the z-score for a 90% confidence interval.
   (b) What is the desired EBP?
   (c) That leaves p', q', and n. For sample size calculation, set $p' = q' = 0.5$. n is unknown, so leave it as n.
(2) Use the EBP formula with the values you have found to solve for n.
(3) Round n = sample size UP to the next whole number.
(4) Write a statement explaining what you have found.

We can use a similar strategy to discover the required sample size when estimating population mean μ. Of course we use the appropriate margin of error formula: $EBM = \text{z-score} * \frac{\sigma}{\sqrt{n}}$.

Example 10. Required Sample Size for a Confidence Interval for μ. Suppose a medical researcher needs to estimate the mean white blood cell count (in cells per microliter) for adults in the US. An analysis of past studies reveals an approximate population standard deviation of 1.6 cells per mL. What sample size is required for a 99% confidence interval of the mean white blood cell count to within 0.2 cells per mL?

Please COMPLETE THE FOLLOWING. **Due: Next class meeting.**

(NOTE: Sample size determination for confidence intervals for population mean μ are not in WebAssign, but you are expected to know it.)

Example 11. How many cars would we need to sample to estimate the average speed of cars passing by Augusta University on Walton Way? Assume we require 95% confidence that the estimate is within 2 miles per hour (mph) of the average speed. A previous study of speeds on a similar road leads us to approximate the standard deviation of speeds at 6 mph.

Example 12. An economist wants to know if gas prices are affecting the proportion of people in the US who commute to work via carpooling. How many people must he sample to be 90% confident that estimate is within 2% of the the true proportion of carpoolers in the US?

CHAPTER 9: HYPOTHESIS TESTING

Lecture Notes for Introductory Statistics [1]

Daphne Skipper, Augusta University (2016)

A **hypothesis test** is a formal way to make a decision based on statistical analysis. A hypothesis test has the following general steps:

- Set up two contradictory hypotheses. One represents our "assumption".
- Perform an experiment to collect data.
- Analyze the data using the appropriate distribution.
- Decide if the experimental data contradicts the assumption or not.
- Translate the decision into a clear, non-technical conclusion.

We will build up all the pieces we need, then put them together into a few complete hypothesis test examples.

1. Null and Alternative Hypotheses.

1.1. **Hypotheses.** Hypothesis tests are tests about a population parameter (μ or p). We will do hypothesis tests about population mean μ and population proportion p.

The **null hypothesis (H_0)** is a statement involving equality ($=, \leq, \geq$) about a population parameter. We assume the null hypothesis is true to do our analysis.

The **alternative hypothesis (H_a)** is a statement that contradicts the null hypothesis. The alternative hypothesis is what we conclude is true if the experimental results lead us to conclude that the null hypothesis (our assumption) is false. The alternative hypothesis must not involve equality ($\neq, <, >$).

The exact statement of the null and alternative hypotheses depend on the claim that you are testing. We take a close look at the steps for writing hypotheses in the following example.

Example 1. Null and Alternative Hypotheses: step-by-step. We wish to determine if the mean time-to-connect in a phone network is less than 3 seconds. Write the null and alternative hypotheses used to test this claim.

(1) Testing a claim about population mean μ or population proportion p?
(2) Use mathematical symbols to express the claim.
(3) Write the opposite of the statement you wrote.
(4) The statement involving equality is the null hypothesis. The other is the alternative. Label them accordingly.
(5) Make a note of the statement you wrote first. This is the claim you need to address in your conclusion.

Example 2. Null and Alternative Hypotheses. We wish to test the claim that the at most 14% of students leave campus for lunch. Write the null and alternative hypotheses for the appropriate hypothesis test.

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

1.2. **Conclusions.** Once you have the null and alternative hypothesis nailed down, there are only two possible **decisions** we can make, based on whether or not the experimental outcome contradicts our assumption (null hypothesis).

(1) Reject the Null Hypothesis (and therefore, Support the Alternative Hypothesis).
(2) Do Not Reject the Null Hypothesis (and therefore, Do Not Support the Alternative Hypothesis).

Since we are basing our conclusion on incomplete information ("sample" data), we must qualify our statements:

(1) The sample data leads us to reject the null hypothesis. (The sample data supports the alternative hypothesis.)
(2) The sample data does not lead us to reject the null hypothesis. (The sample data does not support the alternative hypothesis; it is possible that the null hypothesis is true.)

Note that we can only reject or fail to reject the null hypothesis; we can never "support" the null hypothesis. For this reason, the claim is often represented by the alternative hypothesis, which CAN be supported by a hypothesis test.

Example 3. Stating Conclusions. We wish to determine if the mean time-to-connect in a phone network is less than 3 seconds. Write the two possible conclusions we could draw about this claim using a hypothesis test. (Use the null and alternative hypotheses you found in Example 1.)

(1) Which hypothesis represents the claim?
(2) What is the conclusion about the claim if we decide to reject the null hypothesis?
(3) What is the conclusion about the claim if we decide to fail to reject the null hypothesis?

Example 4. Stating Conclusions. We wish to test the claim that the at most 14% of students leave campus for lunch. Write the two possible conclusions we could draw about this claim using a hypothesis test. (Use the null and alternative hypotheses you found in Example 2.)

(1) Which hypothesis represents the claim?
(2) What is the conclusion about the claim if we decide to reject the null hypothesis?
(3) What is the conclusion about the claim if we decide to fail to reject the null hypothesis?

2. Outcomes and the Type I and Type II Errors

Hypothesis tests are based on incomplete information, since a sample can never give us complete information about a population. Therefore, there is always a chance that our conclusion has been made in error. There are two possible types of error.

The first possible error is if we conclude that the null hypothesis (our assumption) is invalid (choosing to believe the alternative hypothesis), when the null hypothesis is really true. This is called a Type I error.

$$\textbf{Type I error} = \begin{cases} \text{deciding to reject the null when the null is true (RTN)} \\ \text{incorrectly supporting the alternative} \end{cases}$$

The other possible error is if we conclude that the null hypothesis (our assumption) seems reasonable (choosing not to believe the alternative hypothesis), when the null hypothesis is really false. This is called a Type II error.

$$\textbf{Type II error} = \begin{cases} \text{failing to reject the null when the null is False (FRFN)} \\ \text{incorrectly NOT supporting the alternative} \end{cases}$$

It is important to be aware of the probability of getting each type of error. The following notation is used:

$$\alpha = \begin{cases} \text{P(Type I error)} \\ \text{P(decide to reject null} \mid \text{null is true)} \\ \text{P(incorrectly supporting alternative)} \\ \text{the } \textbf{significance level} \text{ of the test} \end{cases}$$

$$\beta = \begin{cases} \text{P(Type II error)} \\ \text{P(decide to “accept” null} \mid \text{null is false)} \\ \text{P(incorrectly NOT supporting alternative)} \end{cases}$$

$$1-\beta = \begin{cases} \text{P(decide to reject null} \mid \text{null is false)} \\ \text{P(correctly supporting the alternative)} \\ \text{the } \textbf{power} \text{ of the test} \end{cases}$$

The **significance level** α is the probability that we *incorrectly* reject the assumption (null) and support the alternative hypothesis. In practice, a data scientist chooses the significance level based on the severity of the consequence of incorrectly supporting the alternative. In our problems, the significance level α will be provided.

The **power** of a test is the probability of correctly supporting a true alternative hypothesis. Usually we are testing if there is statistical evidence to support a claim represented by the alternative hypothesis. The power of the test tells us how often we “get it right” when the claim is true.

Example 5. Error analysis. Suppose the hypotheses are

H_0 : the person being tested does not have HIV, and
H_a : the person does have HIV.

Address each of the following *in the context of this scenario.*

(1) Describe a Type I error.
(2) Describe a Type II error.
(3) Which error is more serious? Explain.
(4) Define α, the significance level.
(5) Define $1-\beta$, the power of the test.
(6) Ideally, α is small and $1-\beta$ is big. In this case, is it more important to minimize α or maximize $1-\beta$?

Example 6. Error analysis. We wish to determine if the mean time-to-connect in a phone network is less than 3 seconds. Address each of the following in the context of this scenario.

(1) Describe a Type I error.
(2) Describe a Type II error.
(3) Which error is more serious for a phone company with connection-time requirements? Explain.

3. Distribution Needed for Hypothesis Testing

The sample statistic (the best point estimate for the population parameter, which we use to decide whether or not to reject the null hypothesis) and distribution for hypothesis tests are basically the same as for confidence intervals.

The only difference is that for hypothesis tests, we assume that the population mean (or population proportion) is known: it is the value supplied by the null hypothesis. (This is how we "assume the null hypothesis is true" when we are testing if our sample data contradicts our assumption.)

When **testing a claim about population mean** μ, ONE of the following two requirements must be met, so that the Central Limit Theorem applies and we can assume the random variable $\bar{X}$ is normally distributed:

- The sample size must be relatively large (many books recommend at least 30 samples), OR
- the sample appears to come from a normally distributed population.

It is very important to verify these requirements in real life. In the problems we are usually told to assume the second condition holds if the sample size is small.

ZTest: The sample statistic is the sample mean of the data, $\bar{x}$. If **population standard deviation is known** (unlikely in real life), the distribution of the sample means is $\bar{X} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$, where μ_0 is the population mean *assumed in the null hypothesis*. The test statistic is a z-score: $z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. The p-value is the tail area under the $\bar{X}$ normal curve beyond $\bar{x}$ in the direction of the alternative hypothesis, which is the same as the tail area under $Z \sim N(0,1)$ beyond z.

TTest: Again the sample statistic is the sample mean of the data, $\bar{x}$. If the **population standard deviation is NOT known**, the distribution of $\bar{X}$ is normally distributed as above, but we use a t-distribution, $t_{df} = t_{n-1}$, since we must use s to approximate σ. The test statistic is a t-score: $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$. The p-value is the area under the T-distribution density curve, t_{df}, beyond t in the direction of the alternative hypothesis.

NOTE: for a two-tailed test (alternative hypothesis is a statement of inequality, $\neq$), the p-value is the tail area outside of the sample statistic *doubled*, which is the probability of getting an outcome this extreme in EITHER direction.

When **testing a claim about population proportion** p, the requirements are that the same as those for a binomial distribution:

- a specified number of indepent trials (n)
- each trial has two outcomes

- each trial has the same probability of "success"

PLUS we require that there are *at least 5 successes and at least 5 failures* ($x = np \geq 5$ and $n - x = nq \geq 5$), so that the binomial distribution has a nice bell shape and can be approximated by a normal distribution.

1-PropZTest: The sample statistic is the sample proportion from the sample data, $p' = \frac{x}{n}$. The distribution of the sample proportions is approximated by a normal distribution with the same mean and standard deviation as the associated binomial distribution: $P' \sim N(p_0, \sqrt{\frac{p_0 q_0}{n}})$, where p_0 is the population proportion *assumed in the null hypothesis*. The test statistic is a z-score: $z = \frac{p' - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$. The p-value is the tail area under the density curve for P' beyond the sample statistic p' in the direction of the null hypothesis, which is the same as the tail area under the standard normal density curve $Z \sim N(0, 1)$ beyond the test statistic z.

4. Rare Events, the Sample, Decision and Conclusion

4.1. **Rare Events.** We decide to reject the null hypothesis if the sample outcome contradicts our assumption. The logic is as follows:

- To make this decision we calculate the

  p-value := the probability of our sample outcome *or something more extreme* occuring ASSUMING the null hypothesis is true.

- If the p-value is very small, our sample outcome was very unlikely (a "rare event") based on our assumption, so we reject our assumption.
- Recall that our assumption is based on the null hypothesis, so rejecting the assumption is the same as rejecting the null hypothesis.
- The smaller the p-value, the larger (in absolute value) the z/t-score, so an "unusual" z/t-score gives us a hint about the p-value.

Example 7. Skittles. Our null hypothesis (assumption) is that a bag of Skittles contains 95% orange Skittles (and the rest are green), because that is written on the label. We reach in a hand and pull out a big handful of Skittles. They are all green!

- Describe the p-value in words.
- We can't get the exact p-value, but what do you estimate it to be?
- Should we decide to "reject" or "not to reject" our assumption (null hypothesis) that the bag contains 95% orange Skittles?

Disclaimer: This scenario is a complete fabrication. Skittles are fabulous.

4.2. **Decision.** How small must the p-value be to reject the null hypothesis? This can change and is set by the **significance level**, α, which will be provided in each problem. Common significance levels are 0.01, or $\frac{1}{100}$, and 0.05, or $\frac{5}{100}$.

The decision process is as follows:

- If p-value $< \alpha$, reject the null hypothesis. (Our outcome contradicts the assumption.)
- If p-value $\geq \alpha$, do not reject the null hypothesis. (Our outcome seems reasonable based on the assumption.)

This silly little rhyme is from the Triola textbook "Elementary Statistics" and has been helpful to many students:

"If the p is low, the null must go.
If the p is high, the null will fly."

5. Additional Information and Full Hypothesis Test Examples.

5.1. **Notes about calculator functions for hypothesis tests.** In practice, we will use the calculator functions TTest, ZTest, and 1-PropZTest to do the calculations for hypothesis tests. You must be able to select the appropriate test, know how to set up the test in the calculator, and interpret the results.

- Use TTest when testing a claim about μ, σ unknown.
- Use ZTest when testing a claim about μ, σ known.
- Use 1-PropZTest when testing a claim about p.
- μ_0 or p_0 should be set to the value of the parameter assumed in the null hypothesis, NOT to the sample value.
- Choose the symbol $\neq$, $<$, or $>$ that corresponds to the alternative hypothesis. Test types:
  - Left-tailed: (H_a is "$<$") p-value is area to the left of the sample statistic. α is the area in the left tail.
  - Right-tailed: (H_a is "$>$") p-value is area to the right of the sample statistic. α is the area in the right tail.
  - Two-tailed: (H_a is "$\neq$") p-value is double the tail-area bounded outside the sample statistic. α is split equally between the two tails.
- When testing claims about proportions, remember to use the decimal form rather than the percentage form.
- All other symbols are as expected.
- Output: p-value, test statistic (t or z), sample statitic ($\bar{x}$ or p')

5.2. **Full Hypothesis Test Examples.**

Example 8. Hypothesis Test 1. The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a population standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marcos mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset significance level $\alpha = 0.05$. Assume the throw distances for footballs are normal.

(1) State the null and alternative hypotheses. (Is this a test about μ or p?)
(2) Find the **sample statistic**, the best point estimate for the population parameter we are testing.
(3) Use the appropriate calculator function to calculate the p-value (probability) and the test statistic (z- or t-score)
(4) In words tell what the p-value represents in this scenario.
(5) At a significance level of $\alpha = 0.05$, what is your decision about the null hypothesis: "reject H_0" or "fail to reject H_0"?
(6) Write your conclusion about the claim in a complete sentence. (Is there statistically significant evidence that the new grip is helping Marco throw farther in general?)

TryIt 9.14.

Example 9. Hypothesis Test 2. It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock will grow at a rate that is different from \$5 per week. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors. Is there evidence that the stock will grow at a rate that is different from \$5 per week? TryIt 9.16.

Example 10. Hypothesis Test 3.

My dog has so many fleas,
They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!

Before his shampoo
I counted 42.
At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out
Use the new shampoo or go without?
(Is there evidence that the new shampoo kills more than 25% of fleas in general?)

Example 9.19.

CHAPTER 10: HYPOTHESIS TESTING WITH TWO SAMPLES

Lecture Notes for Introductory Statistics [1]

Daphne Skipper, Augusta University (2016)

[]Chapter 10Hypothesis Testing with Two Samples

In this chapter our hypothesis tests allow us to compare the means (or proportions) of two different populations using a sample from each population.

For example, we might compare the mean SAT score of a group of students who have taken an SAT prep course to the mean score of a different group of students who haven't taken the course. In this way we could determine if there is statistical evidence that the SAT prep course is effective. We will compare population means using two "independent" samples in sections 10.1 (t distribution) and 10.2 (z distribution).

In section 10.3 we test claims about two independent proportions using "independent" samples.

When testing with two "independent" samples, the hypotheses do not involve comparing a mean or proportion to a specified number, as they do when testing with one sample. Instead they involve comparing two means *to each other*. **Our null hypothesis (assumption) is that the two means (or proportions) of the two underlying populations have some condition involving equality: $\leq, \geq$, or $=$. The alternative is that they are different in some way: $>$, $<$, or $\neq$.**

Example 1. Null and alternative hypotheses for two INDEPENDENT samples. Write the null and alternative hypotheses that would be used to test each claim.

(1) The mean daily high temperature in Augusta is higher than the mean daily high temperature in Savannah. *(We have to have some way to distinguish the two population means. We do this with subscripts: μ_1 and μ_2, but we must keep track of which is "1" and which is "2".)*
(2) The percentage of children who contract the flu after using the nasal vaccine is smaller than the percentage of children who contract the flu after using the placebo.

In section 10.4 we return to testing claims about population means, but this time with "paired data". For example, we might again test the efficacy of the SAT prep course, but this time by collecting "before" and "after" SAT scores from a single group of students. This "paired" sample data is handled differently than the two independent samples in sections 10.1 and 10.2. We will subtract each pair (the two test scores from a single student, for example) to create a new data set of "differences" (after score − before score, for example). We will treat this new (single) data set using the methods of chapter 9 (T-Test or Z-Test).

[1]These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

1. Two Population Means with Unknown Standard Deviations.

In this section we use a t-test to compare the population means of two populations using two *independent* samples. Two **samples are independent** if they come from two distinct populations.

1.1. **Requirements.** The requirements for using a **t-test** for *two independent* simple random samples are:

- the combined sample size is small (< 30) AND
- the data appear to be normally distributed AND
- the sample standard deviation is unknown.

1.2. **Distribution.** We are interested in testing if the mean of one population is greater than the mean of the other. For example, we may test if $\mu_1 > \mu_2$, which is equivalent to $\mu_1 - \mu_2 > 0$. Rather than handling the two distributions $\bar{X}_1$ and $\bar{X}_2$ separately, which would be pretty complicated, we work instead with the distribution of *differences* of sample means: $\bar{X}_1 - \bar{X}_2$. The sample statistic in this case is the best point estimate for $\mu_1 - \mu_2$, which is $\bar{x}_1 - \bar{x}_2$. In this scenario, an outcome that supports the alternative hypothesis (contradicts the null hypothesis) would be if $\bar{x}_1 - \bar{x}_2$ is large, since that would mean $\bar{x}_1$ is much larger than $\bar{x}_2$, running counter to our assumption that $\mu_1 = \mu_2$.

Under the requirements listed above, this new distribution of differences of sample means is also normally distributed with the following distribution:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}\right),$$

where the subscripts indicate which population each statistic or parameter is representing.

Recall that a t-score (or z-score) is has the general formula $\frac{\text{data value} - \text{mean}}{\text{standard deviation}}$, so using this distribution we get:

$$\text{test statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

and the p-value is the area under the normal curve beyond $\bar{x}_1 - \bar{x}_2$ in the direction of the alternative hypothesis.

NOTE: Our null hypothesis assumption is that $\mu_1 = \mu_2$, which means that $\mu_1 - \mu_2 = 0$. This observation puts "0" at the center of our normal distribution. We conclude that we have obtained an "unexpected outcome" if the two sample means are very different from one another.

Since we are using sample standard deviations to approximate population standard deviations with a small sample size, we must use a t-test for error correction. The degrees of freedom calculation is quite complicated. The formula is in the book, but understanding the derivation of the formula is beyond the scope of this course. Instead, the degrees of freedom of the T distribution used in each test will be an output of the calculator function that we use to run the test. You will see that the degrees of freedom does increase with sample size, but not in an easily predictable way.

The calculator function we will use for *testing two means with independent samples* when *sample size is small* (less than 30 combined) and *population standard deviations are uknown*: **2-SampTTest.**

1.3. **Two Sample T-test summary and example.** To summarize:
Requirements: Two *independent* simple random samples of quantitative data. Sample sizes are small (less than 30 combined) and population standard deviations are unknown.
Sample statistic: $\bar{x}_1 - \bar{x}_2$ = the difference in the sample means of our two actual data sets.
Distribution: $\bar{X}_1 - \bar{X}_2$ = the difference in sample means of two randomly selected independent samples from the two populations of interest. $\bar{X}_1 - \bar{X}_2 \sim N\left(0, \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}\right)$. However, we must use the Student T distribution because σ is unknown. The *degrees of freedom* for the t distribution has a complicated formula for two independent samples. However, the degrees of freedom is calculated as a result of running the calculator function 2-SampTTest.
Value assumed in the null hypothesis (center of the normal distribution): $\mu_1 - \mu_2 = 0$. (We assume the two underlying populations have the same mean: $\mu_1 = \mu_2$.)
Hypotheses: Hypotheses have the following form:

$$\begin{array}{lccc} H_0: & \mu_1 & (=, \leq, \geq) & \mu_2 \\ H_a: & \mu_1 & (\neq, >, <) & \mu_2. \end{array}$$

Outcome supporting the alternative hypothesis: When there is a big difference between $\bar{x}_1$ and $\bar{x}_2$ in the direction of the alternative hypothesis.
Test statistics: This is a t-score that tells us how many standard deviations our sample statistic $\bar{x}_1 - \bar{x}_2$ is away from the assumed value of 0: $t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$.
P-value: The probability of getting a difference between sample means this large or larger (direction of the the alternative hypothesis) assuming the difference between the population means is 0: p-value = tail area under the t_{df} curve beyond the test statistic t in the direction of the alternative hypothesis.
Calculator test: 2-SampTTest

Example 2. COMPARE TWO POPULATION MEANS: INDEPENDENT SAMPLES; σ UNKNOWN. A simple random sample of 13 four-cylinder cars is obtained, and the braking distance measured. The mean breaking distance is 137.5 ft and the standard deviation is 5.8 ft. A simple random sample of 12 six-cylinder cars is obtained and the braking distances have a mean of 136.3 ft with a standard deviation of 9.7 ft.

Use a 0.05 significance level to test the claim that the mean braking distance of four-cylinder cars is greater than that of six-cylinder cars. Assume that the samples appear to be somewhat normally distributed

(1) Write the null and alternative hypotheses. Express both using both forms. (For example, $H_a : \mu_1 < \mu_2$ can also be expessed as $H_a : \mu_1 - \mu_2 < 0$.)
(2) Calculate the sample statistic $\bar{x}_1 - \bar{x}_2$. Do you suspect that we will reject the assumption that there is no difference between the population sample means?
(3) Choose the distribution (t or z) and calculator function to use: Are we testing a claim about a mean or proportion? Are σ_1 and σ_2 known or is the combined sample size at least 30?
(4) Use the calculator function **2-SampTTest** to find the test statistic and p-value.
(5) Find the degrees of freedom for the t distribution in this case. (Output of 2-SampTTest.)
(6) What is your decision about the null hypothesis?

(7) State the conclusion about the braking distances of 4 cylinder versus 6 cylinder cars.

2. Two Population Means with Known Standard Deviations.

2.1. **Requirements.** A z-test for two samples (2-SampZTest) is appropriate under the following conditions:

- We are comparing population means using two independent samples.
- The population standard deviations are known OR the combined sample size is greater than 30. (In the latter case, it is safe to use the Z distribution, treating s as σ.)

2.2. **Distribution.** This scenario is just like the last section, except this time we somehow know the population standard deviation for both populations or the combined sample size is at least 30. The distribution of $\bar{X}_1 - \bar{X}_2$ is the same, except the sample standard deviations are replaced by population standard deviations:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right),$$

so that the

$$\text{test statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Of course, we don't need the t-distribution when the population standard deviation is known, so we use **2-SampZTest** to find the test statistic and p-value.

2.3. **Two Sample Z-test summary and example.** To summarize:
Requirements: Two *independent* simple random samples of quantitative data. Sample sizes are large (at least 30 combined) OR population standard deviations are known.
Sample statistic: $\bar{x}_1 - \bar{x}_2$ = the difference in the sample means of our two actual data sets.
Distribution: $\bar{X}_1 - \bar{X}_2$ = the difference in sample means of two randomly selected independent samples from the two populations of interest. $\bar{X}_1 - \bar{X}_2 \sim N\left(0, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right)$.
Value assumed in the null hypothesis (center of the normal distribution): $\mu_1 - \mu_2 = 0$.
Hypotheses: Hypotheses have the following form:

$$\begin{array}{lllll} H_0: & \mu_1 & (=, \leq, \geq) & \mu_2 \\ H_a: & \mu_1 & (\neq, >, <) & \mu_2. \end{array}$$

Outcome supporting the alternative hypothesis: When there is a big difference between $\bar{x}_1$ and $\bar{x}_2$ in the direction of the alternative hypothesis.
Test statistics: This is a z-score that tells us how many standard deviations our sample statistic $\bar{x}_1 - \bar{x}_2$ is away from the assumed value of 0: $z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$.
P-value: The probability of getting a difference between sample means this large or larger (direction of the the alternative hypothesis) assuming the difference between the population means is 0: p-value = tail area under the Z (standard normal) curve beyond the test statistic z in the direction of the alternative hypothesis = tail area under the density curve for $\bar{X}_1 - \bar{X}_2$ beyond the sample statistic $\bar{x}_1 - \bar{x}_2$.

Calculator test: 2-SampZTest

Example 3. COMPARE TWO POPULATION MEANS: INDEPENDENT SAMPLES; σ KNOWN. An interested citizen wanted to know if Democratic U. S. senators are older than Republican U.S. senators, on average. On May 26, 2013, the mean age of 30 randomly selected Republican Senators was 61 years 247 days old (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days old (61.704 years) with a standard deviation of 9.55 years. Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5% level of significance. *(Why is this a z-test instead of a t-test?)* TryIt 9.14.

3. COMPARING TWO INDEPENDENT POPULATION PROPORTIONS.

In this section, we compare the *proportions* of two populations. We use two independent samples, one from each population, and count the number of "successes" in each to arrive at sample proportions for each population. We compare the difference in the two sample proportions $p'_1 - p'_2$ to test if they are different enough to provide evidence that the underlying populations have different proportions.

3.1. **Requirements.** It is easy to see this is a binomial scenario, so the requirements are similar to those for testing a claim about a single population proportion. The requirements for testing a claim about two population proportions using 2-PropZTest are

- The two samples are binomially distributed independent simple random samples.
- There are at least 5 successes ($x = np$) and at least 5 failures ($n - x = nq$) in each of the two samples.

When comparing two population proportions we are, of course, interested in the two sample proportions: $p'_1 = \frac{x_1}{n_1}$ and $p'_2 = \frac{x_2}{n_2}$. Also of interest is the "pooled proportion", which is the proportion of successes in both samples combined:

$$\text{pooled proportion} = p_c = \frac{x_1 + x_2}{n_1 + n_2}.$$

The distribution for the differences of the sample proportions can be approximated by a normal distribution under the conditions stated above,

$$P'_1 - P'_2 \sim N\left(p_1 - p_2, \sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right),$$

which makes the test statistic the z-score,

$$z = \frac{(p'_1 - p'_2) - (p_1 - p_2)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

As in the cases above when comparing sample means, we assume the populations have the same proportion of successes in the null hypothesis, so the center of the associated normal distribuiton is $p_1 - p_2 = 0$.

3.2. **Two sample test of proportions summary and example.** To summarize:
Requirements: Two *independent* simple random samples of binomially distributed data. At least 5 successes and 5 failures in EACH sample.
Sample statistic: $p'_1 - p'_2 =$ the difference between the sample proportions of our two actual data sets: $p'_1 = \frac{x_1}{n_1}$ and $p'_2 = \frac{x_2}{n_2}$.
Value assumed in the null hypothesis (center of the normal distribution): $p_1 - p_2 = 0$. (We assume the populations have the same proportion of successes: $p_1 = p_2$.)
Hypotheses: Hypotheses have the following form:

$$\begin{array}{lccc} H_0: & p_1 & (=, \leq, \geq) & p_2 \\ H_a: & p_1 & (\neq, >, <) & p_2. \end{array}$$

Distribution: $P'_1 - P'_2 =$ the difference between the sample proportions of two randomly selected independent samples from the two populations of interest. $P'_1 - P'_2 \sim N\left(0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$
Outcome that supports the null hypothesis: When there is a big difference between p'_1 and p'_2 in the direction of the alternative hypothesis.
Test statistics: This is a z-score that tells us how many standard deviations our sample statistic $p'_1 - p'_2$ is away from the assumed value of 0: $z = \frac{(p'_1 - p'_2) - 0}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
P-value: The probability of getting a difference between sample proportions this large or larger (in the direction of the the alternative hypothesis) assuming the populations have the same proportion of successes.
Calculator test: 2-PropZTest

Example 4. Compare two population proportions. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1. Complete a hypothesis test at the $\alpha = 0.05$ significance level. Chapter 10 Practice 46-50.

4. Comparing population means: Matched or paired samples.

Here we consider *two sets of data* collected from a *single random sample*: the test scores of a group of students before and after taking a prep course, for example. In this case it makes sense to subtract the two test scores of a single student to find the difference in the two scores. If we do this for every student, we create a new set of data called the differences, d. For each student, we calculate the difference:

$$\begin{array}{ccccc} \text{after score} & - & \text{before score} & = & \text{score difference,} \\ y & - & x & = & d. \end{array}$$

We end up with a *single set of data* to run a hypothesis test on, so we can simply run a T-Test with $n - 1$ degrees of freedom, as in the last chapter.

4.1. **Two *Paired* Samples T-test summary and example.** To summarize:
Requirements: Two *paired* quantitative data sets from a single simple random sample. If the sample size is small (less than 30) the data must be approximately

normal. Often paired tests have small sample size with unknown population standard deviation. If so, we must apply the Student T distribution.

Sample statistic: $\bar{d}$ = the sample mean of the differences of the paired data values. To get $\bar{d}$, subtract each data pair, then find the mean of the differences:

(1) For each subject, calculate $d = y - x$.
(2) Find the mean of the d values, $\bar{d}$.

(Note that for *independent* samples, the sample statitstic is the *difference* of the *sample means*: find the mean of the two samples, then subtract those two values: $\bar{x_1} - \bar{x_2}$.)

Value assumed in the null hypothesis (center of the normal distribution): $\mu_d = 0$ (On average, there is no difference between the two data values collected from each subject).

Hypotheses: Hypotheses have the following form:

$$\begin{array}{lllll} H_0: & \mu_d & (=, \leq, \geq) & 0 \\ H_a: & \mu_d & (\neq, >, <) & 0. \end{array}$$

Distribution: $\bar{D}$ = the sample mean of the differences of paired data values from samples of size n: $\bar{D} \sim N(0, \frac{s_d}{\sqrt{n}})$, but because the population standard deviations are unknown, we use $t_{df} = t_{n-1}$.

Outcome that supports the alternative hypothesis: When $\bar{d}$ is very different from 0, in the direction of the alternative hypothesis.

Test statistics: This is a t-score that tells us how many standard deviations our sample statistic $\bar{d}$ is away from the assumed value of 0: $t = \frac{\bar{d}-0}{\frac{s_d}{\sqrt{n}}}$

P-value: The probability of getting a sample mean of differences this large or larger (direction of the the alternative hypothesis) assuming there is no difference on average between the two data values for each subject ($\mu_d = 0$).

Calculator test: TTest

Example 5. Compare two population means; paired quantitative data. A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table below. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Subject	Before	After
A	6.6	6.8
B	6.5	2.4
C	9.0	7.4
D	10.3	8.5
E	11.3	8.1
F	8.1	6.1
G	6.3	3.4
H	11.6	2.0

Example 10.11.

CHAPTER 11: THE CHI-SQUARE DISTRIBUTION

LECTURE NOTES FOR INTRODUCTORY STATISTICS [1]

Daphne Skipper, Augusta University (2016)

1. THE CHI-SQUARE DISTRIBUTION

In this chapter we explore two types of hypothesis tests that require the **Chi-Square Distribution**, χ^2_{df}.

The Chi-Square distribution has only one parameter: df = degrees of freedom. The degrees of freedom depends on the application, as we will see later. Here are a few facts about the Chi-Square distribution. If $X^2 \sim \chi^2_{df}$ the following are true of X^2:

- X^2 is a continuous random variable
- $X^2 = Z^2 + Z^2 + \cdots + Z^2$; X^2 is the sum of df independent squared standard normal random variables
- data values can't be negative: $x \in [0, \infty)$
- $\mu = df$ (the mean of the Chi-Square distribution is the degrees of freedom!)
- $\sigma = \sqrt{2 * df}$
- X^2 is skewed right
- the mean (df) is just to the right of the peak of the density curve
- when $df > 90$, X^2 is approximately normal

2. GOODNESS OF FIT TEST

We use the goodness of fit test to test if a discrete categorical random variable matches a predetermined "expected" distribution. The hypotheses in a goodness of fit test are

H_0 : the actual distribution fits the expected distribution
H_a : the actual distribution does not fit the expected distribution

REQUIREMENT: In order for a chi-square goodness of fit test to be appropriate, the expected value in each category must be at least 5. It may be possible to combine categories to meet this requirement.

Example 1. FAIR DIE? (PART 1) Suppose we wish to test if a die is weighted. We roll the die 120 times and get the following "observed" results.

Roll	Observed	Expected
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

(1) What is the expected distribution of the 120 die rolls? Complete the table.
(2) Is the requirement for a chi-square goodness of fit test satisfied? Explain.

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

(3) Write the null and alternative hypotheses for a goodness of fit test.
(4) I can see that the rolls didn't come out even. What's the point of completing the test?

Chapter 11 Homework 72.

Our goal is to see if the observed values are close enough to the expected values that the differences could be due to random variation or, alternatively, if the differences great enough that we can conclude that the distribution is not as expected. Therefore, our sample statistic (which is also the test statistic in this case) should provide a measure of how far the from "expected" frequencies the "observed" frequencies are, as a group. The **test statistic** for a goodness of fit test is:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

where O = observed frequency, E = expected frequency, and the sum is taken over all the categories.

Example 2. FAIR DIE? (PART 2) Continuing Example 1, find the test statistic to test if the die is weighted. (Do this by hand using a chart and using the calculator lists.)

The test statistic follows a chi-square distribution, χ^2_{df}, with

$$df = \text{number of categories} - 1.$$

Example 3. FAIR DIE? (PART 3) Continue the same example in which we wish to test if the die is weighted.

(1) Find the distribution of the test statistic.
(2) Find the mean of the distribution.
(3) Find the standard deviation of the distribution.
(4) Sketch the density curve of the distribution.

Studying the test statistic formula, the bigger the differences between the observed and expected frequencies, the larger the test statistic. Since the differences are squared and the expected frequencies are always positive, the test statistic is always 0 or positive. The farther beyond the mean the test statistic is, the more evidence we have that the distribution is not as expected.

The **p-value** is the probability of getting the test statistic or one that is even bigger, which is the area in the right tail of the χ^2_{df} distribution. The calculator function χ^2cdf in the DISTR menu that calculates area under the chi-square distribution,

$$\text{area under the } \chi^2_{df} \text{ density curve between } a \text{ and } b = \chi^2\text{cdf}(a, b, df),$$

so we can calculate

$$\begin{aligned} \text{p-value} &= P(x \geq \text{test statistic} | \text{actual distribution fits expected distribution}) \\ &= \chi^2\text{cdf}(\text{test statistic}, 10^9, df). \end{aligned}$$

Lower p-values indicate test statistics that are farther from the value assumed in the null hypothesis, therefore providing more evidence that the actual distribution does not fit the expected distribution. How low must the p-value be to conclude

there is statistical evidence to support the alternative hypothesis? For this we compare the p-value to the significance level, α. As usual:

If the p is low, the null must go,
if the p is high, the null will fly.

(If p-value $< \alpha$, we reject the null hypothesis. If p-value $\geq \alpha$, we don't have sufficient evidence to reject the null hypothesis.)

Example 4. FAIR DIE? (PART 4) Continue the same example in which we wish to test if the die is weighted.

(1) Shade the area representing the p-value on the χ^2_{df} sketch from Example 3.
(2) Find the p-value for this test.
(3) Should we reject or fail to reject the null hypothesis?
(4) Is there sufficient statistical evidence to conclude that the die is weighted?

Example 5. Conduct a hypothesis test to determine if the actual majors of graduating females fit the expected distribution of their majors. The observed data were collected from 5,000 graduating females. Complete a hypothesis test at the $\alpha = 0.05$ significance level to test if the actual distribution of female students to majors matches the expected distribution.

Major	Expected %	Observed Frequency	Expected Frequency
Arts & Humanities	14.0%	670	
Biological Sciences	8.4%	410	
Business	13.1%	685	
Education	13.0%	650	
Engineering	2.6%	145	
Physical Sciences	2.6%	125	
Professional	18.9%	975	
Social Sciences	13.0%	605	
Techical	0.4%	15	
Other	5.8%	300	
Undecided	8.0%	420	

(1) Find the expected frequencies and complete the table.
(2) Are the requirements for a chi-square goodness of fit test satisfied? Explain and adjust the categories if needed.
(3) Write the null and alternative hypotheses.
(4) What is the distribution?
(5) Find the test statistic.
(6) Find the p-value.
(7) Sketch the density curve. Label the mean, the test statistic, and the p-value.
(8) Is there sufficient evidence to conclude that the distribution of majors is not as expected?

3. Test of Independence

We apply a **test of independence** to see if two characteristics are independent. Again the data are categorical with multiple categories. In fact, the data are grouped accoring to two category types, each with multiple categories. The frequency data are organized in contingency tables with rows representing one category type and columns representing the other category type. In a test for independence, we test whether or not the row categories and column categories are independent of each other.

The null and alternative hypotheses are written out with words and always follow this pattern, although the exact wording will change based on the scenario:

H_0 : the row and column categories are independent
H_a : the row and column categories are dependent

Example 6. Independent? (Part 1) In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year students, and nonstudents. The following table categorizes 839 volunteers according to volunteer type and number of hours worked.

OBSERVED	1-3 hours	4-6 hours	7-9 hours	Row total
Community college students	111	96	48	255
Four-year college students	96	133	61	290
Nonstudents	91	150	53	294
Column total	298	379	162	839

We are interested in whether or not the number of hours worked depends on the volunteer type. List the null and alternative hypotheses for a test for independence:

H_0 :

H_a :

Example 11.6.

We will use the same **test statistic** as for goodness of fit tests:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

where O = observed frequency, E = expected frequency, and the sum is taken over all categories in the table.

We have been given the observed frequencies, but we have to calculate the expected frequencies. To do this, we will need to recall that two events A and B are independent if $P(A \text{ AND } B) = P(A)P(B)$. For example, if I want to find the expected number of volunteers who are community college students (A) AND worked 7-9 hours (B) ASSUMING the number of hours worked is independent of volunteer type, then I can calculate,

$$\begin{aligned} \text{expected number of A AND B volunteers} &= P(A \text{ AND } B) * (\text{n}) \\ &= P(A)P(B) * (\text{n}) \\ &= \tfrac{\text{A row total}}{\text{n}} * \tfrac{\text{B row total}}{\text{n}} * (\text{n}) \\ &= \frac{(\text{number of cc students})(\text{number of 7-9 hr workers})}{\text{total number of volunteers}}, \end{aligned}$$

where n = the total number of volunteers.

The general formula for the expected frequency in row i, column j is

$$E = \frac{(\text{row i total})(\text{column j total})}{n}$$

where n = the sample size.

Example 7. INDEPENDENT? (PART 2) Complete the table of expected values corresponding to Example 6. Expected frequencies assuming number of hours worked is independent of volunteer type:

EXPECTED (assume indep.)	1-3 hours	4-6 hours	7-9 hours	Row total
Community college students				255
Four-year college students				290
Nonstudents				294
Column total	298	379	162	839

REQUIREMENT for test for independence: Each *expected* frequency must be at least 5.

Example 8. INDEPENDENT? (PART 3) Referring to the test for independence we began in Example 6:

(1) Is the requirement for test for independence satisfied? Explain.
(2) Calculate the test statistic using calculator lists. It will be convenient to list all the observed frequencies in L1 and all of the corresponding expected frequencies in L2. Then we can compute the test statistic exactly as we did for the goodness of fit test.

Is it easy to see that big values of the test statistic correspond to big differences between observed frequencies and the frequencies we would expect in the case of independence. Therefore, a big test statistic leads us to support the alternative hypothesis (categories are dependent). But how big is big? Again we will use a p-value and compare it to α to decide whether or not to reject the null hypothesis:

$$\begin{aligned} \text{p-value} < \alpha &\Leftrightarrow \text{BIG test statistic} \\ &\Leftrightarrow \text{reject null (reject independence)} \\ &\Leftrightarrow \text{support alternative (support dependence).} \end{aligned}$$

As we know from the last section, this test statistic follows a chi-square distribution: χ^2_{df}. The degrees of freedom formula for a test for independence is different, however:

$$\text{degrees of freedom } (df) = (\text{number of rows } -1)(\text{number of columns } -1).$$

Again the p-value is the right tail probability in the Chi-square distribution:

$$\text{p-value} = \chi^2\text{cdf}(\text{test statistic}, 10^9, df)$$

Example 9. INDEPENDENT? (PART 4) Complete the same test for independence. Use a significance level of $\alpha = 0.05$.

(1) State the distribution followed by the test statistic.
(2) Find the p-value.
(3) State the conclusion of the test. (Is there statistical evidence that the number of hours worked depends on the volunteer type?)

There is a calculator function that will compute the test statistic and the p-value for a χ^2 test of independence. First enter the matrix of observed values: use MATRIX, EDIT, and select [A]. After entering the table of observed values (not including row and column totals), use STAT, Tests, C: χ^2-Test. As you can see, you have the option of entering a table of expected frequencies. You can do this, but you don't need to. For the case of a test of independence, the expected frequencies will be calculated automatically as the default.

Example 10. INDEPENDENT? (PART 5) Use MATRIX and χ^2-Test on your calculator to find the test statistic and p-value for the test of independence we just completed.

Example 11. TREATING STRESS FRACTURES. With respect to stress fractures in a foot bone, does the success rate of the treatment depend on the treatment method, or do all methods of treatment have basically the same success rate? Use the following data and a significance level of $\alpha = 0.01$ to complete a test of independence.

OBSERVED	Success	Failure	Row total
Surgery	54	12	66
Wt-bearing cast	41	51	92
Non-wt-bearing cast 6 weeks	70	3	73
Non-wt-bearing cast $<$ 6 weeks	17	5	23
Column total	182	71	253

(1) State the null and alternative hypotheses for this test of independence.

(2) Complete the table of expected values assuming the success rate is independent of the treatment method. Use two decimal places of accuracy.

EXPECTED (assuming ind.)	Success	Failure	Row total
Surgery			66
Wt-bearing cast			92
Non-wt-bearing cast 6 weeks			73
Non-wt-bearing cast $<$ 6 weeks			23
Column total	182	71	253

(3) Is the requirement for a test of independence satisfied?
(4) Find the distribution of the test statistic, including the degrees of freedom.
(5) Calculate the test statistic using your preferred method.
(6) Calculate the p-value using your preferred method.
(7) Sketch the density curve, marking and labeling the test statistic and p-value.

(8) What is the outcome of the test for independence? (Can we conclude that the success rate depends on the method of treatment or not?)

(Elementary Statistics, Mario Triola.)

CHAPTER 12: LINEAR REGRESSION AND CORRELATION

Lecture Notes for Introductory Statistics [1]

Daphne Skipper, Augusta University (2016)

In this chapter we explore linear relationships between two sets of paired data. If a scatterplot of paired data shows a linear pattern, we can test for **linear correlation** between the two variables. If there we do find a linear correlation between the variables, we can calculate a **linear regression equation** (a line of best fit) which, given the independent variable, will allow us to predict the value of the dependent variable.

1. Scatterplots

A scatterplot is an easy way to examine the relationship between two variables. Typically, the x-axis represents the independent variable (predictor) and the y-axis represents the dependent variable (predicted).

Example 1. Shoe size versus height. In this example, we will see if there is a linear relationship between men's shoe size, x, and height, y.

(1) Record your own (men's) shoe size, x, and height in inches, y. (Convert a women's shoe size into a men's shoe size by subtracting 1.5. For example, a size 8 women's shoe corresponds to a size 6.5 men's shoe.)
(2) As a class we will record and make a scatterplot of our shoe size and height paired data.
(3) Does there appear to be a linear relationship between shoe size and height?

Not all data is correlated. See Figure 1. Does there appear to be a predictable relationship between height and pulse rate? Another way to think about it: can you predict a person's pulse rate using her height?

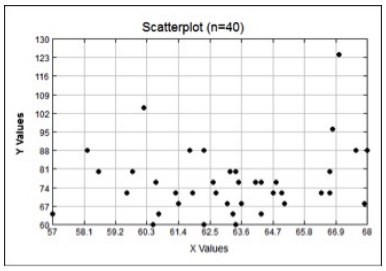

Figure 1. Height (x) versus Pulse rate (y)

Some variable pairs are correlated, but not linearly. The graph in Figure 2 appeared on Dirk Riehle's computers science blog in a paper entitled Software Research and the Industry. There is clearly a pattern present in the data, but it is not linear (it appears to be exponential).

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

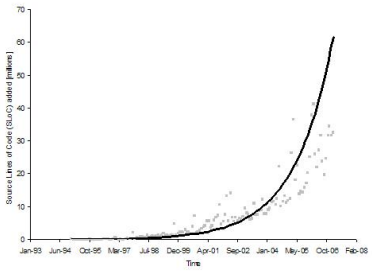

FIGURE 2. Time (x) versus Lines of open source code (y)

2. The Regression Equation

AFTER you have looked at a scatterplot and determined there is linear pattern between two variables, it makes sense to find the line of "best fit". We can use the line to estimate a y value that corresponds to a given x value. The symbol $\hat{y}$ is used to indicate an estimated y-value:

$$\hat{y} = \text{ estimated } y \text{ value.}$$

We need to have a way of quantifying how well the line fits the data, so we consider **residuals**, or **errors**. The residual associated with a data pair (x, y) is the vertical distance between the data point and the approximating line:

$$\epsilon = y - \hat{y} = \text{residual (error).}$$

The residual (error) quantifies how far the estimate $\hat{y}$ (provided by the line) is from the actual data value y. If the point is above the line, the residual is positive. If the point is below the line, the residual is negative.

The **Sum of Squared Errors** is the sum of all the residuals squared (to eliminate negatives):

$$SSE = \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2.$$

The SSE is the measurement we use to determine how well our approximating line fits the data. Smaller errors = better fit.

The **regression equation** is the equation of the line that produces the smallest SSE. The regression equation is also called the **line of best fit** because it minimizes the overall error of the approximations.

BEWARE: A computer (or calculator) can find a linear regression equation for ANY set of data values. However, the regression equation only makes sense if the data actually follow a (somewhat) linear pattern!

Example 2. Shoe size versus height. Use your calculator to graph the scatterplot and the regression equation for our shoe size v. height data.

(1) Enter the shoe sizes into L_1. Enter the heights into L_2.
(2) Make the scatterplot:
    (a) 2^{nd}, Y=
    (b) Plot1: On, Type: (first one), XList: L_1, Ylist: L_2
    (c) ZOOM, 9:ZoomStat
(3) Calculate the regression equation:
    (a) STAT, TESTS, LinRegTTest
    (b) Xlist: L_1, Ylist: L_2, Freq: 1, β & ρ: $\neq 0$
    (c) At RegEQ, choose VARS, Y-VARS, 1:Function, 1:Y_1, ENTER

(d) Calculate

(4) Write the regression equation. (Looking at the output of LinRegTTest, the format of the regression equation is at the top: $\hat{y} = a + bx$. The values of a and b are provided (round these to one more decimal place than the data).
(5) To see the graph of the regression line, select the GRAPH key.
(6) Calculate the residual associated with the data of someone in the class: $y - \hat{y}$. (y is the person's true height. $\hat{y}$ is the estimated height, the output of the regression equation evaluated for $x =$ the person's shoe size.)

One of the results of LinRegTTest is the **correlation coefficient,** r. The correlation coefficient indicates how strong the linear correlation is, and the direction of the correlation. The following are properties of the correlation coefficient, r:

- r is always between -1 and 1.
- The closer r is to 1 or -1, the stronger the linear correlation. If r is close to 0, the correlation is weak. (See Figure 3.)
- If r is negative, the correlation is negative (as x increases, y decreases). This corresponds to a negative slope of the regression equation.
- If r is positive, the correlation is positive (as x increases, y increases). This corresponds to a positive slope of the regression equation. (See Figure 4.)

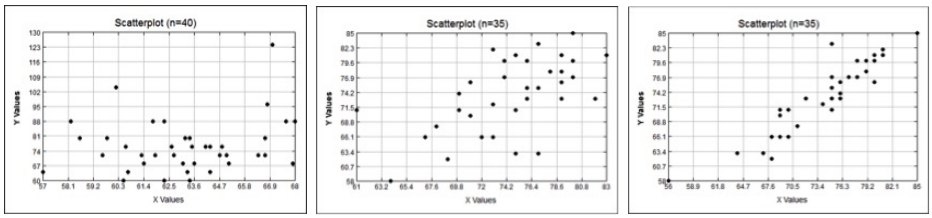

FIGURE 3. $\mathbf{r = 0.202}$ (left), $\mathbf{r = 0.572}$ (center), $\mathbf{r = 0.925}$ (right)

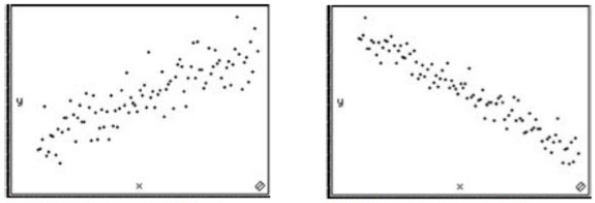

FIGURE 4. Sign of r = sign of slope of regression equation.

The **coefficient of determination**, r^2, usually written as a percentage, has the following significance.

- r^2 represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression line. (A higher r^2 indicates a tighter relationship between x and y.)
- $1 - r^2$ represents the percent of variation in y that is NOT explained by variation in x using the regression line. (The greater $1 - r^2$, the greater the scattering of the observed data points about the regression line.)

Example 3. Shoe size versus height.

(1) Find and interpret the correlation coefficient, r. (Does r indicate positive or negative correlation? Strong or weak correlation?)
(2) Find and interprest the coefficient of determination, r^2.

3. Testing the Significance of the Correlation Coefficient

We know that an r value close to zero indicates low correlation and an r value close to -1 or 1 indicates strong correlation *in the sample data*. What if we want to say something about the underlying population? We can use the following hypothesis test about the ***population* correlation coefficient,** ρ.

Hypothesis test for linear correlation

Hypotheses:

H_0 : $\rho = 0$ (no linear correlation in the population of paired data)

H_a : $\rho \neq 0$ (linear correlation exists in the population of paired data)

Test statistic: Correlation coefficient, r

r critical values, r_c: Use a table to find the $\pm r_c$ critical value corresponding to degrees of freedom, $df = n - 2$ and the significance level ($\alpha = 0.05$ in this book). (The critical values provide a cutoff for how close to 1 or -1 the computed r has to be to imply linear correlation in the underlying population, at a significance level of α.)

p-value: A result of the LinRegTTest. p-value is the probability of getting a correlation coefficient this far from 0 assuming there is no correlation. As usual, a low p-value indicates that we should reject the null hypothesis and support the alternative.

Decision:

$|r| > r_c \Leftrightarrow p - value < \alpha$
$\Leftrightarrow$ Reject H_0
$\Leftrightarrow$ Support H_a
$\Leftrightarrow$ Evidence supports existence of linear correlation in population.

Example 4. Shoe size versus height.

(1) Find the r critical values and draw a number line (from -1 to 1) showing the "reject null" and "don't reject null" regions.
(2) Plot the calculated value of r on the number line to see where it falls.
(3) Find the p-value and compare it to $\alpha = 0.05$.
(4) Is there evidence of a linear correlation between shoe size and height in the population?

4. Prediction

If there is evidence of a linear correlation in the population, it makes sense to use the regression equation to estimate the dependent variable, y, that corresponds to a given value of the independent variable, x. We can refer to x as the *predictor* variable and y as the *predicted* variable.

PREDICTION REQUIREMENTS:

- It makes sense to use the regression equation to predict y ONLY IF there is evidence of a linear correlation AND the scatterplot shows a linear pattern.
- It usually doesn't make sense to predict y for values of x that are outside of the range of data values, since we do not have information about the data outside of this range.

Example 5. Shoe size versus height.

(1) Does it make sense to predict height using shoe size? Explain.
(2) Predict the height of a person who wears a men's size 5.
(3) Predict the height of a person who wears a men's size 10.
(4) Shaquille O'Neal wears a size 22. Does it make sense to predict his height using our regression equation? Explain. Try it anyway then look up his height to check the error.

5. Outliers

Ouliers are data values that have large residuals (errors). Graphically, these are points that are far from the regression line (measured vertically). The standard deviation of the residuals, $s = \sqrt{\frac{SSE}{n-2}}$, is one of the results of the LinRegTTest. Numerically, a data point that has a residual greater than 2s (in absolute value) is considered an outlier.

Example 6. Shoe size versus height.

(1) Find the standard deviation of the residuals, s, in the LinRegTTest output.
(2) Look at the scatterplot with the linear regression line. Do any points appear to be outliers?
(3) Choose a point that looks like it could be an outlier and calculate the residual. Is it an outlier according to the numeric definition of an outlier given here?
(4) If a data value doesn't lie on the regression line, does that mean it is an outlier?

Example 7. The following are advertised sale prices of color televisions at Andersons.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

(1) Decide which variable should be the independent variable and which should be the dependent variable.
(2) Draw a scatter plot of the data.
(3) Does it appear from inspection that there is a relationship between the variables? Why or why not?
(4) Calculate the least-squares line (regression equation). Put the equation in the form of: $\hat{y} = a + bx$.
(5) Find the correlation coefficient. Is it significant? Can we conclude that there is a linear correlation in population? Explain.
(6) Does it make sense to use the regression equation to predict the cost of a 39" television? Explain.

Chapter 12 Homework 85.

Example 8. The table below shows the average heights for American boys in 1990.

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

(1) Decide which variable should be the independent variable and which should be the dependent variable.

(2) Draw a scatter plot of the data.

(3) Does it appear from inspection that there is a relationship between the variables? If so, is it a linear relationship?

(4) Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a+bx$.

(5) Find the correlation coefficient. Is it significant? Can we conclude that there is a linear correlation in population? Explain.

(6) Does it make sense to use the regression equation to predict the average height of a one-year-old boy in 1990? Explain.

(7) Find the estimated average height for a one-year-old.

(8) Find the estimated average height for an eleven-year-old.

(9) Does it appear that a line is the best way to fit the data? Why or why not?

(10) Are there any outliers in the data?

(11) Use the least squares line to estimate the average height for a sixty-two-year-old man. Do you think that your answer is reasonable? Why or why not?

(12) What is the slope of the least-squares (best-fit) line? Interpret the slope.

Chapter 12 Homework #86. (Barbara Illowsky & Susan Dean. Introductory Statistics. OpenStax College, 2013.)

CHAPTER 13: ANALYSIS OF VARIANCE

LECTURE NOTES FOR INTRODUCTORY STATISTICS [1]

Robert Scott and Neal Smith, Augusta University (2016)

In this chapter we explore the problem of determining if, given multiple groups of data, there appear to be any significant mean differences between those groups. A companion problem involves testing whether or not two groups of data have equal variances (standard deviations).

1. One-way ANOVA

A common problem in statistics involves conducting comparisons between a number of different groups. In an earlier chapter, we saw how to compare two groups of data using t-tests, but if more than two groups are under consideration, the analysis is somewhat more complicated, as illustrated in the following example.

Example 1. The wrong way to approach this problem. A university teaches 30 sections of a certain course, and each section takes a common final exam. It is believed that the scores on the exam in each section will be normally distributed and each section will have the same variance (standard deviation). The university wants to determine if there are any differences in mean score on the final among the 30 sections, using $\alpha = .01$ to make these determinations.

An **incorrect** way to approach this problem would be to begin by testing all possible pairs of sections against one another using t-tests. Since $\alpha = .01$, this would mean that each test carries a 1 percent probability of a type I error. We would have to perform a total of $\frac{30\times 29}{2} = 435$ t-tests to accomplish this, as there are 30 choices for the first group to consider, and since we will not test a group against itself, 29 choices for the second group; however this counts each pair twice (for example we test A versus B a second time as B versus A), hence the need to divide by 2. There is a 99 percent probability we do not make a type I error on each test, and so the probability we would make no type I errors in this process is $.99^{435}$, which is only about 1.3 percent. Therefore, there is a 98.7 percent probability we will make one or more type I errors during the testing process (where two sections that had the same mean performance are determined to be different). So, we will end up doing a bunch of work and it's almost certain that all this work will contain some type I errors, which doesn't sound like a good idea.

Therefore, we would like to do a **single test** to determine if any significant differences exist among these 30 sections. Such a problem is called a **One-Way ANOVA** (Analysis of Variance).

There are a number of assumptions we must make to carry out an ANOVA.

(1) Each population from which a sample is taken is assumed to be normal.
(2) All samples are randomly selected and independent.
(3) The populations are assumed to have equal standard deviations (and therefore equal variances).
(4) The null hypothesis to be tested is that all the populations have equal means.

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

(5) The alternate hypothesis is that one or more pairs of the populations have unequal means.

Example 2. In Psychology, there is a well-documented phenomenon known as the **anchoring effect** where the first piece of information received by a subject tends to greatly influence their perception. A student wishes to test the anchoring effect by taking a large jar which contains 1000 pennies and asking a number of subjects to visually estimate the number of pennies in the jar. However, before doing so, the experimenter hands each subject a large card with a number written on it-either 500, 1000, or 1500. The data below gives each subject's estimate of the number of pennies in the jar. We will assume that the guesses in each of the three groups are normally distributed with equal variance.

Group A (shown 500 card)	Group B (shown 1000 card)	Group C (shown 1500 card)
500	1330	1270
480	1150	1470
840	720	1360
470	1130	1360
660	1220	1410
500	890	1150
730	760	1130
910	1020	1640
540	950	1330

First, we wish to determine if there is any mean difference between any of the three groups, and let's use $\alpha = .05$ to make a conclusion. First, enter this data as three separate lists on the TI-83/84 (say lists L_1, L_2, and L_3). Then under STAT-TESTS, use the ANOVA command (the syntax will be ANOVA(L_1, L_2, L_3), as the arguments in the parentheses are the sets of data we wish to compare).

In the output, F denotes the value of the test statistic. We will not discuss the specifics of calculating the test statistic, but under the ANOVA assumptions, the test statistic can be shown to have something called an F-distribution. We are most interested in the p-value, which in this case is 4.89×10^{-8}. Since the p-value is less than α, we reject the null hypothesis and conclude that two or more of these three groups have a different mean.

Now that we have determined that there are some differences among the three groups, we will now test each of the three groups against each other, but our first example shows that without a little care, we need to worry about making a type I error. There are a number of ways to do this, but the simplest way is to use a smaller alpha when doing each of these individual tests. In fact, if we divide our original alpha of .05 by the number of tests to be done (in this case, three: the 500 group vs. the 1000 group, the 500 vs. the 1500, and the 1000 vs. the 1500), it can be shown that this will control the type I error rate appropriately. So for each of the individual t-tests we will use a significance level 0f $.05/3 \approx .0167$.

One the calculator, when we use the two sample t-test (we are comparing normally distributed populations and we must estimate the standard deviation from sample data), the calculator will ask us if we want to **pool** the data. This is appropriate if each group has the same standard deviation, which we are assuming in this problem. Doing so, we find the following p-values.

(1) 500 group vs. 1000 group two-sided p-value 4.2×10^{-4}
(2) 500 group vs. 1500 group two-sided p-value 6.1×10^{-8}

(3) 1000 group vs. 1500 group two-sided p-value .0016

Since each of these are less than our adjusted α of .05/3, we conclude that each of these groups have a different mean, or in other words, the card that the subject was shown seems to have a significant effect on the number of pennies they reported in the jar.

Example 3. Thirty-two similar individuals who are showing the initial symptoms of a common cold are given one of 4 treatments-either a placebo or one of three brands of over the counter cold remedy. The patients are then tracked and the time for their cold symptoms to resolve in days is noted. The data is below.

Placebo	Brand A	Brand B	Brand C
13	8	6	3
11	6	10	7
9	8	9	5
10	10	9	4
10	8	8	3
12	11	7	5
12	7	9	8
11	5	6	7

We want to carry out an ANOVA to see if there are any significant differences in the mean time needed for the cold symptoms to resolve between the four treatment groups, using $\alpha = .05$ to make the conclusion. As before, we will assume all the ANOVA requirements are satisfied.

Check that the p-value of the ANOVA is 4.1×10^{-6}, so we reject the null hypothesis that each of the four groups has the same mean. Now that we have a reason to go looking for differences in the four groups, also check that the two-sided p-values are as follows.

(1) Placebo vs. treatment A two-sided p-value .0022
(2) Placebo vs. treatment B two-sided p-value 8.2×10^{-4}
(3) Placebo vs. treatment C two-sided p-value 6.0×10^{-6}
(4) Treatment A vs. treatment B two-sided p-value .89
(5) Treatment A vs. treatment C two-sided p-value .017
(6) Treatment B vs. treatment C two-sided p-value .0065

Since we are doing 6 comparisons, a significance level of $.05/6 \approx .0083$ can be used for the individual comparisons. We see that there is a statistically significant difference in the mean time needed for the symptoms to resolve in each of the treatments compared to the placebo. Treatment C also seems different than treatment B, but the p-values on A versus B and A versus C means that A seems significantly different from only the placebo.

2. Testing Standard Deviations

Recall that the variance is defined to be the square of the standard deviation; in comparing variances, we are also comparing standard deviations. The variance is what is described as an unbiased estimator whereas the standard deviation is a biased estimator; that means that we can use assumptions of normality and similar mathematical constraints in studying the variance directly, but not in studying the

standard deviations. We will make conclusions about the standard deviations only indirectly: by finding out about variances.

The purpose of this test is to look at data from two samples and think about the variance of the two underlying populations. Are their variances (and hence their standard deviations) statistically the same? This is important in several settings. For example, if two machines are manufacturing equivalent repair parts for your automobile and the parts from the two machines average the same size (their means are equal), but one machine produces parts with a larger variance than the other, you would favor the other machine. Why? Because the more consistently sized parts will work better when the engine is repaired!

In order for this test to be mathematically accurate:

(1) Each population from which a sample must be normally distributed. This criterion is more significant for the Test of Two Variances than it was for earlier tests we have studied.
(2) All samples are randomly selected and the two populations are independent of each other (there may not even be the same sample size from the two populations).

Example 4. Two statistics classes of 35 students each took a test, and it was found the average time needed for the students to finish was very similar in each class. However, in the first class, the sample standard deviation of the time needed to finish was 7 minutes, and in the second section the sample standard deviation was 8.5 minutes. Therefore, we would like to know if the variance in finishing times for the first class was significantly less than for the second class. Let's use $\alpha = .05$ to make a conclusion.

We are testing the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ versus the alternate $H_1 : \sigma_1^2 < \sigma_2^2$; that is we initially assume that the two groups have equal variances, but we could potentially be convinced that the variance of class one was significantly smaller.

To measure whether we see a statistically significant effect, it turns out that the correct way to proceed is to use the test statistic $F = \frac{(s_1)^2}{(s_2)^2}$, and under our assumptions, this ratio will have a so-called F-distribution. For our data, $F = \frac{7^2}{8.5^2} = .6782$. Now, if class one had a smaller variance than class two, we would expect the value of the F statistic to be small, and so to compute the p-value, we would want to consider the probability $P(F < .6782)$. To find this probability on the TI-83/84, under the DISTR menu, we could select the Fcdf command (cumulative probability with an F distribution) and input $Fcdf(0, .6782, 34, 34)$. To dissect the syntax, we want the probability that F is between 0 and .6782 (note that F must be a non-negative value), and the last two arguments mean that our two samples each have 35-1 = 34 degrees of freedom as discussed in chapter 11. We then find that the p-value is .1313, and since this p-value is not less than alpha, we do not reject the null and therefore we do not conclude that there was any significant difference in the variance of the finishing times of these two classes.

We could also have computed a test statistic of $F = \frac{(s_2)^2}{(s_1)^2} = \frac{8.5^2}{7^2} = 1.4745$, but if class one has a smaller variance, the numerator of F will be much larger than the denominator, and therefore F would tend to be fairly large. Therefore, the p-value of the test set up in this way would be $Fcdf(1.4745, 10000, 34, 34)$, which is .1313 just as before.

SUPPLEMENTARY TOPIC: CHEBYSHEV'S THEOREM

Lecture Notes for Introductory Statistics [1]

Neal Smith, Augusta University (2016)

In chapter 2, we discussed how to calculate the mean and standard deviation of a data set. As it turns out, simply knowing the mean and standard deviation of a data set can reveal much about the nature of a set of data. A well-known result attributed to Chebyshev makes the nature of this statement precise.

1. Chebyshev's Theorem

First, we begin by stating Chebyshev's Theorem. We will not prove the result.

Chebyshev's Theorem. Given any set of data, and any real number $k \geq 1$, at least $1-1/k^2$ of the points in that set of data must fall within k standard deviations of the mean. That is, at least $1-1/k^2$ of the data must lie in the interval between $\mu - k\sigma$ and $\mu + k\sigma$.

This is a pretty powerful result, since it allows us to make a statement about any set of data. Additionally, there are some special cases of this theorem that are worth knowing, so let's take a look at a couple of these.

If we set $k = 2$, then $1 - 1/k^2 = 1 - 1/4 = 75\%$, and Chebyshev's Theorem tells us that in any data set, at least 75 percent of the data must lie within $k = 2$ standard deviations of the mean.

If we set $k = 3$, then $1 - 1/k^2 = 1 - 1/9 \approx 88.9\%$, and Chebyshev's Theorem tells us that in any data set, at least 88.9 percent of the data must lie within $k = 3$ standard deviations of the mean.

If we set $k = 4$, then $1 - 1/k^2 = 1 - 1/16 = 93.75\%$, and Chebyshev's Theorem tells us that in any data set, at least 93.75 percent of the data must lie within $k = 4$ standard deviations of the mean.

We could of course do this all day, but the takeaway is that in any set of data we can expect the vast majority of the data to fall within just a few standard deviations of the mean.

2. Examples Using Chebyshev's Theorem

Example 1. A group of 20 students were asked for the amount they spent in textbooks during the last academic year. The amounts in dollars reported were

700, 600, 550, 550, 550, 500, 500, 500, 450, 450,
450, 400, 400, 400, 400, 350, 350, 300, 300, 200

Check that this population has mean 445 dollars with $\sigma = 113.91$ dollars.

Chebyshev's Theorem therefore predicts that at least 75 percent of the costs will fall in the interval $445 - 2 \times 113.91 = 217.18$ to $445 + 2 \times 113.91 = 672.82$. By examination of the data set, we see that this is certainly true, as 18 of the 20 (or 90 percent) of the reported costs do fall in this interval. Thus, the operative phrase is **at least**.

[1] These lecture notes are intended to be used with the open source textbook "Introductory Statistics" by Barbara Illowsky and Susan Dean (OpenStax College, 2013).

Also observe that Chebyshev's Theorem predicts that at least 88.9 percent of the costs fall within three standard deviations of the mean, but it is easy to see that in fact all of the data in this particular data set is in fact within three standard deviations of the mean.

Example 2. A professor tells a class that the mean on a recent exam was 80 with a standard deviation of 6 points, and suppose you wanted to find an interval where at least 75 percent of the students must have scored. Since 75 percent corresponds to $k = 2$ in Chebyshev's Theorem, we need only look 2 standard deviations from the mean to conclude that at least 75 percent of the students scored between $80 - 2 \times 6 = 68$ and $80 + 2 \times 8 = 96$.

Depending on the data, Chebyshev's Theorem may tell you a lot or not so much. Let's look at an example where Chebyshev's Theorem is not too enlightening.

Example 3. A professor tells a class that the mean on a recent (100 point) exam was 62 and the standard deviation was a whopping 33 points. Again, if we wanted to get a handle on at least 75 percent of the exam scores, we would let $k = 2$ and conclude that at least 75 percent of the students scored between $62 - 2 \times 33 = -4$ and $62 + 2 \times 33 = 128$. Since this was a 100 point exam, and presumably negative scores were not possible, this tells us that at least 75 percent of the students scored between 0 and 100 on the exam. While this statement is of course true, it is not terribly enlightening! Hopefully, you can see that since the standard deviation was so large, this is an indication of high variability in the data set, and there is simply too much potential variation in the data to be able to draw fantastic conclusions knowing only the mean and the standard deviation!

Let's do one final example.

Example 4. Anew college graduate has done their homework and is searching for their first job. Based on their major, their educational level, the type of job they are looking for, their experience, and the geographic location where they want to live, a salary aggregator tells them that the mean salary of new employees is approximately 45000 dollars with a standard deviation of 2600 dollars. This person is subsequently offered a salary of 52000 dollars. How good is this offer?

Solution. Well, it's certainly not terrible, being above the mean, but fortunately we can quantify this somewhat better. First, since Chebyshev's Theorem can tell us what is happening a certain number of standard deviations away from the mean, it would be nice to know a z-score for this 52000 dollar salary.

$$z_{52000} = \frac{x - \mu}{\sigma} = \frac{52000 - 45000}{2600} \approx 2.7$$

Since this salary is 2.7 standard deviations away from the mean, using Chebyshev's Theorem with $k = 2.7$ tells us that at least $1 - 1/2.7^2 \approx 86.3$ percent of the salaries must lie within 2.7 standard deviations of the mean; that is between 38000 and 52000 dollars. Thus, we can safely conclude that at least this 52000 dollar offer is greater or equal to at least 86.3 percent of the other salaries out there.